AFRL-RY-WP-TR-2013-0083

THE EXPONENTIALLY EMBEDDED FAMILY OF DISTRIBUTIONS FOR EFFECTIVE DATA REPRESENTATION, INFORMATION EXTRACTION, AND DECISION MAKING

Steven Kay, Haibo He, and Quan Ding

University of Rhode Island

MARCH 2013
Final Report

Approved for public release; distribution unlimited

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals.

Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RY-WP-TR-2013-0083 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

*//Signature//

MURALIDHAR RANGASWAMY, Program Manager
Radio Frequency Exploitation Technology

//Signature//

JEFFREY SANDERS, Branch Chief
Radio Frequency Exploitation Technology

//Signature//

DOUG HAGER, Deputy
Layered Sensing Exploitation Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

*Disseminated copies will show "//Signature//" stamped or typed above the signature blocks.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS**.

1. REPORT DATE ***(DD-MM-YY)*** March 2013	**2. REPORT TYPE** Final	**3. DATES COVERED *(From - To*)** 20 September 2011 – 20 December 2012

4. TITLE AND SUBTITLE

THE EXPONENTIALLY EMBEDDED FAMILY OF DISTRIBUTIONS FOR EFFECTIVE DATA REPRESENTATION, INFORMATION EXTRACTION, AND DECISION MAKING

5a. CONTRACT NUMBER
FA8650-11-1-7148

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61101E

6. AUTHOR(S)

Steven Kay, Haibo He, and Quan Ding

5d. PROJECT NUMBER
1000

5e. TASK NUMBER
11

5f. WORK UNIT NUMBER
Y000

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Rhode Island
Kingston, RI 02881

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-RY-WP-TR-2013-0083

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Sensors Directorate
Wright-Patterson Air Force Base, OH 45433-7320
Air Force Materiel Command
United States Air Force

Defense Advanced Research Projects
Agency (DARPA/DSO)
675 North Randolph Street
Arlington, VA 22203-2114

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RYAP

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RY-WP-TR-2013-0083

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

Report contains color.

14. ABSTRACT

We have focused on the mathematical formalism and stochastic machine learning algorithms for extraction of relevant information based on the exponentially embedded family (EEF), learning and classification over large-scale stream data, and information fusion and integration. In particular, we have proposed a probability density function (PDF) estimation approach based on the EEF, and a measure for assessment of information from sensors. We have also taken advantage ofthe model structure information for model estimation. Furthermore, we have proved a general Pythagorean theorem for the EEF and studied a multi path scenario for sensor selection. Finally, we also analyzed and developed a series of machine learning techniques for effective data learning, classification, and decision making, including adaptive incremental learning from stream data, information fusion with multiple learning models/hypotheses, machine learning with non-stationary imbalanced stream data, kernel density estimation based on self-organizing map(SOM), among others. These results have been published in peer-reviewed conferences and journals, including IEEE Transactions on Neural Networks, IEEE Transactions on Neural Networks and Learning Systems, Neurocomputing (Elsevier), a book chapter with Wiley-IEEE, among others.

15. SUBJECT TERMS

information extraction, incremental learning, classification, decision making

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON (Monitor)**
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	SAR	42	Muralidhar Rangaswamy **19b. TELEPHONE NUMBER *(Include Area Code)*** N/A

1 Summary

The extraction of relevant information from large-scale data requires an accurate model of the data source. In practice, such a model has to be generic enough to encompass data sources of different modality but also mathematically tractable to allow a fast rate of learning. We have selected exponentially embedded family (EEF) of probability density functions (PDFs) for information extraction. The EEF belongs to the exponentially family of PDFs, and therefore inherits many important properties to allow efficient information extraction and learning from data. First of all, it admits sufficient statistics and therefore, provides the means for selecting good models. It also easily allows additional sensor statistics to be evaluated and possibly incorporated into the model. The exponential family is a reproducible probability density function family (a Gaussian is a special case), whose conjugate density is also of the exponential family type. For unsupervised scenarios in which unknown parameters need to be estimated, the maximization of the likelihood leads to a convex optimization problem, and thus can be easily implemented. Furthermore, the model is sufficiently general to encompass any real-world situation. The embedded exponential family approach yields a means of on-line assessment of performance since the Kullback-Liebler divergence is a part of the model. The rate of learning is also readily found since the Kullback-Liebler divergence can be used to ascertain distances between PDFs for various hypothesis testing scenarios.

We have focused on the mathematical formalism and stochastic machine learning algorithms for extraction of relevant information based on the EEF, learning and classification over large-scale stream data, and information fusion and integration. In particular, we have proposed a PDF estimation approach based on the EEF, and a measure for assessment of information from sensors. We have also taken advantage of the model structure information for model estimation. Furthermore, we have proved a general Pythagorean theorem for the EEF and studied a multipath scenario for sensor selection. Finally, we also analyzed and developed a series of machine learning techniques for effective data learning, classification, and decision making, including adaptive incremental learning from stream data, information fusion with multiple learning models/hypotheses, machine learning with non-stationary imbalanced stream data, kernel density estimation based on self-organizing map(SOM), among others. These results have been published in peer-reviewed conferences and journals, including IEEE Transactions on Neural Networks, IEEE Transactions on Neural Networks and Learning Systems, Neurocomputing (Elsevier),

a book chapter with Wiley-IEEE, among others.

2 EEF for Estimation of PDF

Consider a problem where we have two sensors (extension is straightforward for multiple sensors) as shown in Figure 1. We use the EEF to assess the significance of the information contributed by T_2 for a decision.

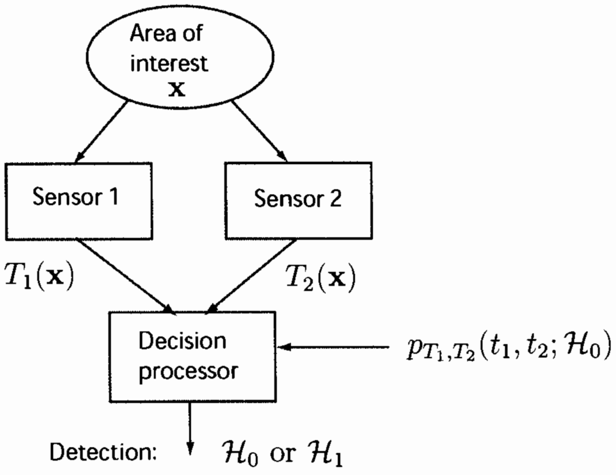


Figure 1: Signal Detection Problem Overview

1. Assume PDF for background ($\mathcal{H}_0$) is known - in practice can measure this before event of interest takes place.
2. Construct best approximation to PDF using PDF under $\mathcal{H}_0$ and T_1 and T_2 data under $\mathcal{H}_1$.
3. Produces EEF PDF approximation

$$p_{\boldsymbol{\eta}}(\mathbf{t}) = \exp\left[\eta_1 t_1 + \eta_2 t_2 - K(\eta_1,\eta_2) + \ln p_T(\mathbf{t};\mathcal{H}_0)\right]$$

We have to choose the *embedding parameters* η_1, η_2. Note that if we decide $\eta_2 = 0$, this is equivalent to ignoring sensor 2 output.

To minimize the Kullback-Liebler distance (*determines prob. of detection performance*), we choose embedding parameters so that approximate PDF has moments *matched* to true PDF under $\mathcal{H}_1$, which are presumed known, i.e.,

$$E_t[T_1] = \lambda_1 \qquad E_t[T_2] = \lambda_2 \qquad \text{(from data under } \mathcal{H}_1)$$

This is equivalent to Gram-Schmidt orthogonalization for Gaussian PDFs (see Figure 2).

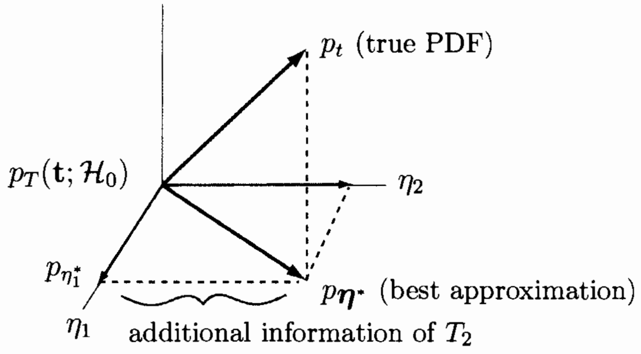

p_t (true PDF)
$p_T(\mathbf{t};\mathcal{H}_0)$
η_2
$p_{\eta_1^*}$
$p_{\boldsymbol{\eta}^*}$ (best approximation)
η_1
additional information of T_2

Figure 2: Best Approximation

For one sensor we construct

$$p_{\eta_1^*} = p_{\eta_1^*,\eta_2=0}(t_1,t_2)$$

and for two sensors we construct

$$p_{\boldsymbol{\eta}^*} = p_{\eta_1^*,\eta_2^*}(t_1,t_2)$$

Information content of $T_2(\mathbf{x})$ is

$$D(p_{\eta_1^*,\eta_2^*}(t_1,t_2)||p_{\eta_1^*,\eta_2=0}(t_1,t_2)) = \text{ reduction in distance to true PDF}$$

where $D(p_1||p_2)$ is Kullback-Liebler distance

$$D(p_1||p_2) = \int p_1(\mathbf{x}) \ln \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} d\mathbf{x}$$

3 Assessment of Information from Sensors for the EEF

We have proposed a measure of information increase for the exponentially embedded family, when new sensors are added. This measure is always greater than or equal to zero, which implies that by adding new sensors, we could always obtain some information (or at least keep the same information if new sensors are redundant). We have proved that the information provided by independent sensors is additive. Based on this measure, we can decide which sensor provides the most information and select the best combination of sensors. We can also find redundant sensors if this measure becomes zero. Therefore, we can perform sensor selection and sensor reduction using this measure of information increase.

Assume that we have only one sensor $\mathbf{T}_1(\mathbf{x})$, then the EEF is constructed as

$$p_{\boldsymbol{\eta}_1}(\mathbf{x}) = \exp\left[\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x}) - K_1(\boldsymbol{\eta}_1) + \ln p_0(\mathbf{x})\right] \tag{1}$$

where $K_1(\boldsymbol{\eta}_1) = \ln\left(\int \exp\left[\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x})\right]p_0(\mathbf{x})d\mathbf{x}\right) = \ln E_0\left(\exp\left[\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x})\right]\right)$. Since $\boldsymbol{\eta}_1$ are the unknown parameters, we find the MLE of $\boldsymbol{\eta}_1$ by maximizing $p_{\boldsymbol{\eta}_1}(\mathbf{x})$ or equivalently maximizing $\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x}) - K_1(\boldsymbol{\eta}_1)$. Taking the derivative and setting it to zero, the MLE should satisfy

$$\mathbf{T}_1(\mathbf{x}) = \left.\frac{\partial K_1(\boldsymbol{\eta}_1)}{\partial \boldsymbol{\eta}_1}\right|_{\hat{\boldsymbol{\eta}}_1} \tag{2}$$

The KL divergence between the true PDF $p_t(\mathbf{x})$ and the EEF $p_{\boldsymbol{\eta}_1}(\mathbf{x})$ is

$$\begin{aligned}
D(p_t||p_{\boldsymbol{\eta}_1}) &= \int p_t(\mathbf{x}) \ln \frac{p_t(\mathbf{x})}{p_{\boldsymbol{\eta}_1}(\mathbf{x})} d\mathbf{x} \\
&= \int p_t(\mathbf{x}) \left[\ln p_t(\mathbf{x}) - \ln p_0(\mathbf{x}) - \left(\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x}) - K_1(\boldsymbol{\eta}_1)\right)\right] d\mathbf{x} \\
&= D(p_t||p_0) - \left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) - K_1(\boldsymbol{\eta}_1)\right]
\end{aligned} \tag{3}$$

Let $\boldsymbol{\eta}_1^{(1)*}$ be the $\boldsymbol{\eta}_1$ that minimizes $D(p_t||p_{\boldsymbol{\eta}_1})$ or equivalently maximizes $\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) - K_1(\boldsymbol{\eta}_1)$. Then $\boldsymbol{\eta}_1^{(1)*}$ should satisfy

$$E_t\left(\mathbf{T}_1(\mathbf{x})\right) = \left.\frac{\partial K_1(\boldsymbol{\eta}_1)}{\partial \boldsymbol{\eta}_1}\right|_{\boldsymbol{\eta}_1^{(1)*}} \tag{4}$$

In the case when we have L IID unobserved samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ and we only observe L IID outputs $\mathbf{T}_1(\mathbf{x}_1), \mathbf{T}_1(\mathbf{x}_2), \ldots, \mathbf{T}_1(\mathbf{x}_L)$ from the same sensor, (2) can be extended as

$$\frac{1}{L}\sum_{i=1}^{L} \mathbf{T}_1(\mathbf{x}_i) = \left.\frac{\partial K_1(\boldsymbol{\eta}_1)}{\partial \boldsymbol{\eta}_1}\right|_{\hat{\boldsymbol{\eta}}_1} \tag{5}$$

Since $\frac{1}{L}\sum_{i=1}^{L} \mathbf{T}_1(\mathbf{x}_i) \xrightarrow{P} E_t\left(\mathbf{T}_1(\mathbf{x})\right)$ as $L \to \infty$, it can be shown that

$$\hat{\boldsymbol{\eta}}_1 \xrightarrow{P} \boldsymbol{\eta}_1^{(1)*} \tag{6}$$

and

$$D(p_t||p_{\hat{\boldsymbol{\eta}}_1}) \xrightarrow{P} D(p_t||p_{\boldsymbol{\eta}_1^{(1)*}}) = D(p_t||p_0) - \max_{\boldsymbol{\eta}_1} \left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) - K_1(\boldsymbol{\eta}_1)\right] \quad (7)$$

Say we add another sensor $\mathbf{T}_2(\mathbf{x})$, the EEF becomes

$$p_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(\mathbf{x}) = \exp\left[\boldsymbol{\eta}_1^T \mathbf{T}_1(\mathbf{x}) + \boldsymbol{\eta}_2^T \mathbf{T}_2(\mathbf{x}) - K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) + \ln p_0(\mathbf{x})\right] \quad (8)$$

where $K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \ln E_0\left(\exp\left[\boldsymbol{\eta}_1^T \mathbf{T}_1(\mathbf{x}) + \boldsymbol{\eta}_2^T \mathbf{T}_2(\mathbf{x})\right]\right)$. Note that when $\boldsymbol{\eta}_2 = \mathbf{0}$, the EEF in (10) reduces to the EEF in (1) because

$$K(\boldsymbol{\eta}_1, \mathbf{0}) = \ln E_0\left(\exp\left[\boldsymbol{\eta}_1^T \mathbf{T}_1(\mathbf{x})\right]\right) = K_1(\boldsymbol{\eta}_1) \quad (9)$$

The MLE of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are obtained by solving

$$\begin{aligned} \mathbf{T}_1(\mathbf{x}) &= \left.\frac{\partial K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{\partial \boldsymbol{\eta}_1}\right|_{\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2} \\ \mathbf{T}_2(\mathbf{x}) &= \left.\frac{\partial K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{\partial \boldsymbol{\eta}_2}\right|_{\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2} \end{aligned} \quad (10)$$

The KL divergence between the true PDF $p_t(\mathbf{x})$and the EEF $p_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}(\mathbf{x})$ is

$$D(p_t||p_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}) = D(p_t||p_0) - \left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\right] \quad (11)$$

Let $\boldsymbol{\eta}_1^{(2)*}$ and $\boldsymbol{\eta}_2^{(2)*}$ be the $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ that minimize $D(p_t||p_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2})$ or equivalently maximize $\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. Then $\boldsymbol{\eta}_1^{(2)*}$ and $\boldsymbol{\eta}_2^{(2)*}$ satisfy

$$\begin{aligned} E_t\left(\mathbf{T}_1(\mathbf{x})\right) &= \left.\frac{\partial K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{\partial \boldsymbol{\eta}_1}\right|_{\boldsymbol{\eta}_1^{(2)*}, \boldsymbol{\eta}_2^{(2)*}} \\ E_t\left(\mathbf{T}_2(\mathbf{x})\right) &= \left.\frac{\partial K(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{\partial \boldsymbol{\eta}_2}\right|_{\boldsymbol{\eta}_1^{(2)*}, \boldsymbol{\eta}_2^{(2)*}} \end{aligned} \quad (12)$$

Similarly, it can be shown that

$$\begin{aligned} \hat{\boldsymbol{\eta}}_1 &\xrightarrow{P} \boldsymbol{\eta}_1^{(2)*} \\ \hat{\boldsymbol{\eta}}_2 &\xrightarrow{P} \boldsymbol{\eta}_2^{(2)*} \end{aligned}$$

and

$$\begin{aligned}
&D(p_t||p_{\hat{\boldsymbol{\eta}}_1,\hat{\boldsymbol{\eta}}_2}) \xrightarrow{P} D(p_t||p_{\boldsymbol{\eta}_1^{(2)*},\boldsymbol{\eta}_2^{(2)*}}) \\
&= D(p_t||p_0) - \max_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\right]
\end{aligned}$$

Since

$$\begin{aligned}
&\max_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\right] \\
&\geq \max_{\boldsymbol{\eta}_1}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\big|_{\boldsymbol{\eta}_2=\mathbf{0}}\right] \\
&= \max_{\boldsymbol{\eta}_1}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) - K_1(\boldsymbol{\eta}_1)\right] \qquad (13)
\end{aligned}$$

where the last equality is from (11). Therefore, from (9) and (14), we conclude that KL divergence between the true PDF and the EEF will decrease by adding another sensor. The difference

$$\begin{aligned}
&D(p_t||p_{\boldsymbol{\eta}_1^{(1)*}}) - D(p_t||p_{\boldsymbol{\eta}_1^{(2)*},\boldsymbol{\eta}_2^{(2)*}}) \\
&= \max_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\right] - \max_{\boldsymbol{\eta}_1}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) - K_1(\boldsymbol{\eta}_1)\right] \\
&\qquad (14)
\end{aligned}$$

can be considered as a measure of the information increase by adding another sensor.

Also note that if $\mathbf{T}_1(\mathbf{x})$ and $\mathbf{T}_2(\mathbf{x})$ are independent under $\mathcal{H}_0$, then

$$\begin{aligned}
K(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) &= \ln E_0\left(\exp\left[\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x}) + \boldsymbol{\eta}_2^T\mathbf{T}_2(\mathbf{x})\right]\right) \\
&= \ln\left[E_0\left(\exp\left[\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x})\right]\right) E_0\left(\exp\left[\boldsymbol{\eta}_2^T\mathbf{T}_2(\mathbf{x})\right]\right)\right] \\
&= \ln E_0\left(\exp\left[\boldsymbol{\eta}_1^T\mathbf{T}_1(\mathbf{x})\right]\right) + \ln E_0\left(\exp\left[\boldsymbol{\eta}_2^T\mathbf{T}_2(\mathbf{x})\right]\right) \\
&= K_1(\boldsymbol{\eta}_1) + K_2(\boldsymbol{\eta}_2) \qquad (15)
\end{aligned}$$

where $K_2(\boldsymbol{\eta}_2)$ is the normalizing factor or cumulant generating function of the EEF

$$p_{\boldsymbol{\eta}_2}(\mathbf{x}) = \exp\left[\boldsymbol{\eta}_2^T\mathbf{T}_2(\mathbf{x}) - K_2(\boldsymbol{\eta}_2) + \ln p_0(\mathbf{x})\right] \qquad (16)$$

constructed using the sensor $\mathbf{T}_2(\mathbf{x})$ only. In this case,

$$\begin{aligned}
&\max_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\right] \\
&= \max_{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) + \boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K_1(\boldsymbol{\eta}_1) - K_2(\boldsymbol{\eta}_2)\right] \\
&= \max_{\boldsymbol{\eta}_1}\left[\boldsymbol{\eta}_1^T E_t\left(\mathbf{T}_1(\mathbf{x})\right) - K_1(\boldsymbol{\eta}_1)\right] + \max_{\boldsymbol{\eta}_2}\left[\boldsymbol{\eta}_2^T E_t\left(\mathbf{T}_2(\mathbf{x})\right) - K_2(\boldsymbol{\eta}_2)\right] \\
&\qquad (17)
\end{aligned}$$

and (17) becomes

$$\begin{aligned} & D(p_t || p_{\boldsymbol{\eta}_1^{(1)*}}) - D(p_t || p_{\boldsymbol{\eta}_1^{(2)*}, \boldsymbol{\eta}_2^{(2)*}}) \\ &= \max_{\boldsymbol{\eta}_2} \left[\boldsymbol{\eta}_2^T E_t \left(\mathbf{T}_2(\mathbf{x}) \right) - K_2(\boldsymbol{\eta}_2) \right] \end{aligned} \tag{18}$$

4 Model Estimation via Model Structure Determination

In model estimation, we often face problems with unknown parameters in the candidate models. This paper proposes the model structure determination (MSD) for model estimation with unknown parameters. We start with the problem of model order selection, and decompose the probability density function (PDF) into the information provided by the data about the model parameters and that of the model structure. The factor that depends on the model parameters is approximated using a minimax procedure, and the MSD depends on the model structure only. It is shown that the MSD is equivalent to the exponentially embedded family (EEF) for model order selection under some conditions. Finally, we apply the MSD to a classification problem where we have partial knowledge about the parameters, and simulation results show that it outperforms the pseudo-maximum likelihood (pseudo-ML) rule.

4.1 Model Order Structure

We start by considering the problem of model order selection, where we have a set of candidate models $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M\}$. Each model has a set of unknown parameters which we denote as $\boldsymbol{\theta}_i$ and which has dimension $p_i \times 1$. Based on an observation of $\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$, we wish to choose a model *without knowledge of* $\boldsymbol{\theta}_i$ for each model. The PDF is given as $p_X(\mathbf{x}; \boldsymbol{\theta}_i, \mathcal{M}_i)$. Furthermore, we assume that a minimal sufficient statistic exists for $\boldsymbol{\theta}_i$ and is given by $\mathbf{T}_i(\mathbf{x})$. Dispensing with the particular model notation for the time being, note that by the Neyman-Fisher factorization theorem the PDF can be written as

$$p_X(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{t}(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}) \tag{19}$$

where $\mathbf{t}(\mathbf{x})$ is the $p \times 1$ sufficient statistic for $\boldsymbol{\theta}$, which is also $p \times 1$. We next decompose the PDF into the information provided by the data about the *model* *parameters* and that of the *model structure*. To do so we have from (19) that

$$\frac{p_X(\mathbf{x};\boldsymbol{\theta})}{p_X(\mathbf{x};\boldsymbol{\theta}_0)} = \frac{g(\mathbf{t}(\mathbf{x}),\boldsymbol{\theta})}{g(\mathbf{t}(\mathbf{x}),\boldsymbol{\theta}_0)} \tag{20}$$

where $\boldsymbol{\theta}_0$ is arbitrary. Next assume that $g(\mathbf{t}(\mathbf{x}),\boldsymbol{\theta})$ can be normalized to integrate to one or if $\mathbf{t} = \mathbf{t}(\mathbf{x})$, then

$$\int g(\mathbf{t},\boldsymbol{\theta})d\mathbf{t} = c < \infty$$

for c a constant, which does not depend on $\boldsymbol{\theta}$. As a result, the PDF of $\mathbf{t}(\mathbf{x})$ is

$$p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}) = g(\mathbf{t}(\mathbf{x}),\boldsymbol{\theta})/c$$

and we can write (20) as

$$p_X(\mathbf{x};\boldsymbol{\theta}) = p_X(\mathbf{x};\boldsymbol{\theta}_0)\frac{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)}$$

and finally this becomes

$$p_X(\mathbf{x};\boldsymbol{\theta}) = \underbrace{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})}_{\text{model parameters}} \quad \underbrace{\frac{p_X(\mathbf{x};\boldsymbol{\theta}_0)}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)}}_{\text{model structure}} . \tag{21}$$

The first factor depends only on the *model parameters* and since $\boldsymbol{\theta}_0$ is arbitrary the second factor depends only on the *model structure* (it is functionally independent of $\boldsymbol{\theta}$), i.e., the dimensionality of the model. Note that in the conditional model estimator (CME), only the second factor is used by omitting the first factor. Here, we use both factors. To illustrate the ideas we will use the linear model, which is important in practice.

For the linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{H}$ is $N \times p$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$ with σ^2 known, it is well known that the minimal sufficient statistic is $\mathbf{t}(\mathbf{x}) = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ and furthermore that $\mathbf{t}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\theta},\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. Thus,

$$p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2}|\sigma^2(\mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{t}(\mathbf{x})-\boldsymbol{\theta})^T\frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}(\mathbf{t}(\mathbf{x})-\boldsymbol{\theta})\right]$$

and by writing

$$(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}) = (\mathbf{x}-\mathbf{H}\mathbf{t}(\mathbf{x}))^T(\mathbf{x}-\mathbf{H}\mathbf{t}(\mathbf{x}))+(\mathbf{t}(\mathbf{x})-\boldsymbol{\theta})^T(\mathbf{H}^T\mathbf{H})(\mathbf{t}(\mathbf{x})-\boldsymbol{\theta})$$

we have that

$$\begin{aligned} p_X(\mathbf{x};\boldsymbol{\theta}) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\mathbf{t}(\mathbf{x}))^T(\mathbf{x}-\mathbf{H}\mathbf{t}(\mathbf{x}))\right] \\ &\quad \cdot \exp\left[-\frac{1}{2}(\mathbf{t}(\mathbf{x})-\boldsymbol{\theta})^T\frac{(\mathbf{H}^T\mathbf{H})}{\sigma^2}(\mathbf{t}(\mathbf{x})-\boldsymbol{\theta})\right]. \end{aligned}$$

As a result the model structure component of the PDF is

$$\begin{aligned} \frac{p_X(\mathbf{x};\boldsymbol{\theta}_0)}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)} &= \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\mathbf{t}(\mathbf{x}))^T(\mathbf{x}-\mathbf{H}\mathbf{t}(\mathbf{x}))\right]}{\frac{1}{(2\pi)^{p/2}|\sigma^2(\mathbf{H}^T\mathbf{H})^{-1}|^{1/2}}} \\ &= \frac{1}{(2\pi\sigma^2)^{(N-p)/2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{P}_H^{\perp}\mathbf{x}\right] \underbrace{\frac{1}{|\mathbf{H}^T\mathbf{H}|^{1/2}}}_{\text{Jacobian}}. \qquad (22) \end{aligned}$$

Note that $\mathbf{P}_H^{\perp} = \mathbf{I}-\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ (the orthogonal projector operator) annihilates the signal $\mathbf{H}\boldsymbol{\theta}$ and hence $\mathbf{x}^T\mathbf{P}_H^{\perp}\mathbf{x}$ does not depend on $\boldsymbol{\theta}$. This again shows that the model structure component is functionally independent of $\boldsymbol{\theta}$. When $\mathbf{x}$ is transformed to $\mathbf{t}$ and $\mathbf{u}$, then the Jacobian is needed. More specifically, let the $N \times (N-p)$ matrix $\mathbf{B} = [\mathbf{b}_1\,\mathbf{b}_2 \ldots \mathbf{b}_{N-p}]$ consist of columns that are orthonormal and span the orthogonal subspace to the columns of $\mathbf{H}$. Hence, we have that

$$\mathbf{x} = \mathbf{H}\mathbf{t} + \mathbf{B}\mathbf{u} \qquad (23)$$

where $\mathbf{u}$ is $(N-p)\times 1$. Also, we note that $\mathbf{B}^T\mathbf{B} = \mathbf{I}_{N-p}$ and $\mathbf{B}^T\mathbf{H} = \mathbf{0}$. Hence, the transformation from $\mathbf{x}$ to $(\mathbf{t},\mathbf{u})$ is from (23)

$$\mathbf{x} = \underbrace{[\mathbf{H}\,\mathbf{B}]}_{\mathbf{A}} \begin{bmatrix} \mathbf{t} \\ \mathbf{u} \end{bmatrix}.$$

The Jacobian is

$$|\mathbf{A}^T\mathbf{A}|^{1/2} = \left| \begin{bmatrix} \mathbf{H}^T \\ \mathbf{B}^T \end{bmatrix} [\mathbf{H}\,\mathbf{B}] \right|^{1/2} = \left| \begin{bmatrix} \mathbf{H}^T\mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^T\mathbf{B} = \mathbf{I}_{N-p} \end{bmatrix} \right|^{1/2} = |\mathbf{H}^T\mathbf{H}|^{1/2}$$

which cancels the Jacobian term in (22). Also,

$$\mathbf{x}^T\mathbf{P}_H^{\perp}\mathbf{x} = \left([\mathbf{H}\,\mathbf{B}]\begin{bmatrix} \mathbf{t} \\ \mathbf{u} \end{bmatrix}\right)^T \mathbf{P}_H^{\perp} \left([\mathbf{H}\,\mathbf{B}]\begin{bmatrix} \mathbf{t} \\ \mathbf{u} \end{bmatrix}\right)$$

and noting that $\mathbf{P}_H^\perp \mathbf{H} = \mathbf{0}$ and $\mathbf{P}_H^\perp \mathbf{B} = \mathbf{B}$, we have that

$$\mathbf{x}^T \mathbf{P}_H^\perp \mathbf{x} = \begin{bmatrix} \mathbf{t} \\ \mathbf{u} \end{bmatrix}^T \underbrace{\begin{bmatrix} \mathbf{H}^T \\ \mathbf{B}^T \end{bmatrix} [\mathbf{0}\,\mathbf{B}]}_{\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N-p} \end{bmatrix}} \begin{bmatrix} \mathbf{t} \\ \mathbf{u} \end{bmatrix} = \mathbf{u}^T \mathbf{u}.$$

Also, since $\mathbf{t}$ is a complete sufficient statistic and $\mathbf{u}$ does not depend on $\boldsymbol{\theta}$, and hence is ancillary, by Basu's theorem $\mathbf{t}$ and $\mathbf{u}$ are independent. Hence, we have finally from (21) and (22) after transformation to $\mathbf{t}$ and $\mathbf{u}$ that

$$\begin{aligned} \mathbf{t} &\sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}) \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N-p}) \end{aligned} \tag{24}$$

and are independent.

For inference on $\boldsymbol{\theta}$ we would generally discard $\mathbf{u}$ and make decisions based on the sufficient statistic. This assumes that we wish to test *within a given PDF family.* However, when choosing between models, i.e., between PDF families, it is just the opposite. The sufficient statistic, indicating information about the model parameters, is of no use (when they are unknown). It is actually $\mathbf{u}$ that is important in choosing between models. The main difficulty in just discarding $\mathbf{t}$ is that the remaining data vector $\mathbf{u}$ changes in dimension as the model changes and so any decision is based on various dimensionality data sets. This will lead to unacceptable results. In other words, we cannot simply compare $p_{U_i}(\mathbf{u}_i|\mathcal{M}_i)$'s for model estimation since $\mathbf{u}_i$'s may have different dimensions. We must maintain the dimensionality of the data by replacing the PDF of $\mathbf{t}$ by one *that is known.* We next show how to do this.

We keep the part of the PDF of $(\mathbf{t}, \mathbf{u})$ that is associated with $\mathbf{u}$ and attempt to replace the part associated with $\mathbf{t}$ since without knowledge of $\boldsymbol{\theta}$, it is of no use for model estimation as discussed at the end of the previous section. Hence, the problem now reduces to finding a suitable approximation or estimate of $p_T(\mathbf{t}; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}) = \mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)$. In order to arrive at a meaningful solution we will need to constrain the space of $\boldsymbol{\theta}$. Otherwise, our approach is not viable, as will be shown later. Hence, we assume that $\boldsymbol{\theta}^T \mathbf{C}_t^{-1} \boldsymbol{\theta} = \frac{\boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}}{\sigma^2} \leq \xi^2$, which is the interior and boundary of an ellipsoid in R^N, so that the possible values of $\boldsymbol{\theta}$ lie within $\Theta = \{\boldsymbol{\theta} : \frac{\boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}}{\sigma^2} \leq \xi^2\}$. In many cases of model estimation we may already have some idea as to the possible limits of the parameters. For example, Θ can be considered as a weighted energy constraint which is often encountered in practice such as communications. Hence, such a constraint may also make practical sense. With this restriction we let the approximating PDF be $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ and choose $\boldsymbol{\mu}$ and $\mathbf{C}$ to best approximate the original PDF $\mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. To do so we adopt the Kullback-Liebler (KL) measure between PDFs and use a minimax approach. We will see that it admits a simple and very intuitive solution and one that under some conditions *coincides with the EEF for model order selection.* The KL measure (also called the divergence) is defined as

$$D(p_T(\mathbf{t};\boldsymbol{\theta})||\hat{p}_T(\mathbf{t};\boldsymbol{\mu},\mathbf{C})) = \int p_T(\mathbf{t};\boldsymbol{\theta}) \ln \frac{p_T(\mathbf{t};\boldsymbol{\theta})}{\hat{p}_T(\mathbf{t};\boldsymbol{\mu},\mathbf{C})} d\mathbf{t}. \tag{25}$$

For the case of two multivariate Gaussian PDFs with $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$ and $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ it is shown to be

$$D(\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{C}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\mathbf{C}_2)) = \frac{1}{2}\ln\frac{|\mathbf{C}_2|}{|\mathbf{C}_1|} + \frac{1}{2}\mathrm{tr}\left[\mathbf{C}_1(\mathbf{C}_2^{-1} - \mathbf{C}_1^{-1})\right] + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{C}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{26}$$

The next theorem shows that the minimax approximating PDF to $\mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$ is given as $\mathcal{N}(\mathbf{0}, (1+\xi^2/p)\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. It is interesting to note that the same result is obtained in the Bayesian case if we were to assume Zellner's *g-prior* for $\boldsymbol{\theta}$. In this case the prior PDF for $\boldsymbol{\theta}$ is $\mathcal{N}(\mathbf{0}, (\xi^2/p)\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. This is undoubtedly due to the close connection between minimax and Bayesian decision theories.

Theorem 1 (Minimax approximation to PDF). *Assume a* $p\times 1$ *random vector* $\mathbf{T}$ *is distributed according to* $p_T = \mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)$ *where the mean* $\boldsymbol{\theta}$ *is unknown but lies within a constraint set* $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\theta}^T\mathbf{C}_t^{-1}\boldsymbol{\theta} \le \xi^2\}$*, and the covariance matrix* $\mathbf{C}_t$ *is known. We approximate the PDF of* $\mathbf{T}$ *in a minimax sense using* $\hat{p}_T = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ *for some* $\boldsymbol{\mu}, \mathbf{C}$*, which is equivalent to solving the problem*

$$\min_{\boldsymbol{\mu},\mathbf{C}} \max_{\boldsymbol{\theta}\in\Theta} D(\mathcal{N}(\boldsymbol{\theta},\mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu},\mathbf{C})) \tag{27}$$

The solution of (27) is $\boldsymbol{\mu}^* = \mathbf{0}$*,* $\mathbf{C}^* = \left(1 + \frac{\xi^2}{p}\right)\mathbf{C}_t$*.*

Proof. *See Appendix A.*

Having obtained a suitable replacement for the PDF of the sufficient statistic of $\boldsymbol{\theta}$, which does not depend on $\boldsymbol{\theta}$, we can finally obtain the approximating PDF for the original data $\mathbf{x}$. From (24) we now have

$$\begin{aligned} \mathbf{t} &\sim \mathcal{N}(\mathbf{0}, (1+\xi^2/p)\sigma^2(\mathbf{H}^T\mathbf{H})^{-1}) \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{N-p}) \end{aligned}$$

and noting that $\mathbf{P}_H^\perp \mathbf{H} = \mathbf{0}$ and $\mathbf{P}_H^\perp \mathbf{B} = \mathbf{B}$, we have that

$$\mathbf{x}^T \mathbf{P}_H^\perp \mathbf{x} = \begin{bmatrix} \mathbf{t} \\ \mathbf{u} \end{bmatrix}^T \underbrace{\begin{bmatrix} \mathbf{H}^T \\ \mathbf{B}^T \end{bmatrix} [\mathbf{0}\,\mathbf{B}]}_{\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N-p} \end{bmatrix}} \begin{bmatrix} \mathbf{t} \\ \mathbf{u} \end{bmatrix} = \mathbf{u}^T \mathbf{u}.$$

Also, since $\mathbf{t}$ is a complete sufficient statistic and $\mathbf{u}$ does not depend on $\boldsymbol{\theta}$, and hence is ancillary, by Basu's theorem $\mathbf{t}$ and $\mathbf{u}$ are independent. Hence, we have finally from (21) and (22) after transformation to $\mathbf{t}$ and $\mathbf{u}$ that

$$\begin{aligned} \mathbf{t} &\sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}) \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N-p}) \end{aligned} \tag{24}$$

and are independent.

For inference on $\boldsymbol{\theta}$ we would generally discard $\mathbf{u}$ and make decisions based on the sufficient statistic. This assumes that we wish to test *within a given PDF family.* However, when choosing between models, i.e., between PDF families, it is just the opposite. The sufficient statistic, indicating information about the model parameters, is of no use (when they are unknown). It is actually $\mathbf{u}$ that is important in choosing between models. The main difficulty in just discarding $\mathbf{t}$ is that the remaining data vector $\mathbf{u}$ changes in dimension as the model changes and so any decision is based on various dimensionality data sets. This will lead to unacceptable results. In other words, we cannot simply compare $p_{U_i}(\mathbf{u}_i|\mathcal{M}_i)$'s for model estimation since $\mathbf{u}_i$'s may have different dimensions. We must maintain the dimensionality of the data by replacing the PDF of $\mathbf{t}$ by one *that is known.* We next show how to do this.

We keep the part of the PDF of $(\mathbf{t}, \mathbf{u})$ that is associated with $\mathbf{u}$ and attempt to replace the part associated with $\mathbf{t}$ since without knowledge of $\boldsymbol{\theta}$, it is of no use for model estimation as discussed at the end of the previous section. Hence, the problem now reduces to finding a suitable approximation or estimate of $p_T(\mathbf{t}; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}) = \mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)$. In order to arrive at a meaningful solution we will need to constrain the space of $\boldsymbol{\theta}$. Otherwise, our approach is not viable, as will be shown later. Hence, we assume that $\boldsymbol{\theta}^T \mathbf{C}_t^{-1} \boldsymbol{\theta} = \frac{\boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}}{\sigma^2} \leq \xi^2$, which is the interior and boundary of an ellipsoid in R^N, so that the possible values of $\boldsymbol{\theta}$ lie within $\Theta = \{\boldsymbol{\theta} : \frac{\boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}}{\sigma^2} \leq \xi^2\}$. In many cases of model estimation we may already have some idea as to the possible limits of the parameters. For example, Θ can be considered as a weighted energy constraint which is often encountered in practice such as communications. Hence, such a constraint may also make practical sense. With this restriction we let the approximating PDF be $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ and choose $\boldsymbol{\mu}$ and $\mathbf{C}$ to best approximate the original PDF $\mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. To do so we adopt the Kullback-Liebler (KL) measure between PDFs and use a minimax approach. We will see that it admits a simple and very intuitive solution and one that under some conditions *coincides with the EEF for model order selection.* The KL measure (also called the divergence) is defined as

$$D(p_T(\mathbf{t};\boldsymbol{\theta})||\hat{p}_T(\mathbf{t};\boldsymbol{\mu},\mathbf{C})) = \int p_T(\mathbf{t};\boldsymbol{\theta}) \ln \frac{p_T(\mathbf{t};\boldsymbol{\theta})}{\hat{p}_T(\mathbf{t};\boldsymbol{\mu},\mathbf{C})} d\mathbf{t}. \tag{25}$$

For the case of two multivariate Gaussian PDFs with $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$ and $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ it is shown to be

$$D(\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{C}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\mathbf{C}_2)) = \frac{1}{2}\ln\frac{|\mathbf{C}_2|}{|\mathbf{C}_1|} + \frac{1}{2}\mathrm{tr}\left[\mathbf{C}_1(\mathbf{C}_2^{-1} - \mathbf{C}_1^{-1})\right] + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{C}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{26}$$

The next theorem shows that the minimax approximating PDF to $\mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$ is given as $\mathcal{N}(\mathbf{0}, (1+\xi^2/p)\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. It is interesting to note that the same result is obtained in the Bayesian case if we were to assume Zellner's *g-prior* for $\boldsymbol{\theta}$. In this case the prior PDF for $\boldsymbol{\theta}$ is $\mathcal{N}(\mathbf{0}, (\xi^2/p)\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. This is undoubtedly due to the close connection between minimax and Bayesian decision theories.

Theorem 1 (Minimax approximation to PDF). *Assume a $p \times 1$ random vector $\mathbf{T}$ is distributed according to $p_T = \mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)$ where the mean $\boldsymbol{\theta}$ is unknown but lies within a constraint set $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\theta}^T\mathbf{C}_t^{-1}\boldsymbol{\theta} \leq \xi^2\}$, and the covariance matrix $\mathbf{C}_t$ is known. We approximate the PDF of $\mathbf{T}$ in a minimax sense using $\hat{p}_T = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ for some $\boldsymbol{\mu}, \mathbf{C}$, which is equivalent to solving the problem*

$$\min_{\boldsymbol{\mu},\mathbf{C}} \max_{\boldsymbol{\theta}\in\Theta} D(\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})) \tag{27}$$

The solution of (27) is $\boldsymbol{\mu}^* = \mathbf{0}$, $\mathbf{C}^* = \left(1 + \frac{\xi^2}{p}\right)\mathbf{C}_t$.

Proof. *See Appendix A.*

Having obtained a suitable replacement for the PDF of the sufficient statistic of $\boldsymbol{\theta}$, which does not depend on $\boldsymbol{\theta}$, we can finally obtain the approximating PDF for the original data $\mathbf{x}$. From (24) we now have

$$\begin{aligned} \mathbf{t} &\sim \mathcal{N}(\mathbf{0}, (1+\xi^2/p)\sigma^2(\mathbf{H}^T\mathbf{H})^{-1}) \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{N-p}) \end{aligned}$$

where $\mathbf{t}$ and $\mathbf{u}$ are independent, and upon transforming back to $\mathbf{x}$ via $\mathbf{x} = \mathbf{H}\mathbf{t}+\mathbf{B}\mathbf{u}$ we have that $\mathbf{x}$ is multivariate Gaussian with a zero mean vector and covariance matrix

$$\begin{aligned}\mathbf{C} &= \mathbf{H}\mathbf{C}_t\mathbf{H}^T + \mathbf{B}\mathbf{C}_u\mathbf{B}^T \\ &= \mathbf{H}(1+\xi^2/p)\sigma^2(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T + \mathbf{B}\sigma^2\mathbf{I}_{N-p}\mathbf{B}^T \\ &= (1+\xi^2/p)\sigma^2\mathbf{P}_H + \sigma^2\mathbf{B}\mathbf{B}^T.\end{aligned}$$

But note that $\mathbf{B}\mathbf{B}^T = \sum_{i=1}^{N-p}\mathbf{b}_i\mathbf{b}_i^T$ is the orthogonal projection operator and so

$$\mathbf{B}\mathbf{B}^T = \mathbf{P}_H^{\perp}.$$

The covariance matrix becomes

$$\mathbf{C} = (1+\xi^2/p)\sigma^2\mathbf{P}_H+\sigma^2\mathbf{P}_H^{\perp} = (1+\xi^2/p)\sigma^2\mathbf{P}_H+\sigma^2(\mathbf{I}_N-\mathbf{P}_H) = \sigma^2\mathbf{I}_N+(\xi^2/p)\sigma^2\mathbf{P}_H$$

and is seen to be an "inflated" covariance. With the above analysis, we conclude this section with the following theorem:

Theorem 2 (Minimax PDF of $\mathbf{x}$ for the linear model). *For the linear model, the minimax PDF of* $\mathbf{x}$ *is* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N + (\xi^2/p)\sigma^2\mathbf{P}_H)$, *where the condition* $\frac{\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}}{\sigma^2} \leq \xi^2$ *determines* ξ^2.

We then consider a signal duration estimation problem and compare the performance of the MSD, EEF and MDL. Consider a model estimation problem with

$$\mathcal{M}_p: \quad x[n] = \begin{cases} s[n]+w[n] & \text{for } n = 0,1,\ldots,p-1 \\ w[n] & \text{for } n = p,p+1,\ldots,N-1 \end{cases}$$

where $s[n]$'s are unknown but satisfy $\sum_{n=0}^{p-1}s^2[n]/\sigma^2 \leq p\xi^2 = \xi_p^2$ with known ξ^2, and $w[n]$ is the white Gaussian noise (WGN) with known variance σ^2. Here, the constraint can be written as $\sum_{n=0}^{p-1}s^2[n]/p \leq \sigma^2\xi^2$, which is a power constraint on the signal and makes physical sense. Note that for $\mathcal{M}_p$, the signal length is p, and therefore, this is a signal duration estimation problem. It can be written as the linear model with

$$\mathbf{H}_p = \begin{bmatrix} \mathbf{I}_p \\ \mathbf{0}_{(N-p)\times p} \end{bmatrix}$$

Approved for public release; distribution unlimited.

and $\boldsymbol{\theta}_p = [s[0]\ s[1]\ \ldots\ s[p-1]]^T$ with constraint

$$||\mathbf{H}_p\boldsymbol{\theta}_p||^2/\sigma^2 = \sum_{n=0}^{p-1} s^2[n]/\sigma^2 \leq p\xi^2$$

The MSD criterion can be simplified as

$$MSD(p) = \frac{1}{2\sigma^2}\frac{\xi^2}{1+\xi^2}\sum_{n=0}^{p-1} x^2[n] - \frac{p}{2}\ln\left(1+\xi^2\right)$$

Let the number of samples be $N = 10$, the largest candidate model order be $M = 3$, and each model has equal prior probabilities. We fix $\sigma^2\xi^2 = 1$ so that the size of the constraint balls is fixed. We plot the probability of correct selection P_C versus $1/\sigma^2$ in Figure 3 (note that for different σ^2, ξ^2 is different, since $\sigma^2\xi^2$ is fixed). For each σ^2, the experiment is repeated 50000 times. For each run, $\mathbf{s}_p$ is uniform randomly distributed within $\sum_{n=0}^{p-1} s^2[n] \leq p\sigma^2\xi^2 = p$. The EEF and the MDL are also simulated for comparison. The EEF and MDL rules choose the model order that maximizes the following respectively:

$$EEF(p) = \begin{cases} 2\ln L_{G_p}(\mathbf{x}) - p\left[\ln\left(2\ln L_{G_p}(\mathbf{x})/p\right) + 1\right] & 2\ln L_{G_p}(\mathbf{x}) \geq p \\ 0 & 2\ln L_{G_p}(\mathbf{x}) < p \end{cases}$$

$$-MDL(p) = 2\ln L_{G_p}(\mathbf{x}) - p\ln N$$

for $p = 1, 2, \ldots, M$, where $2\ln L_{G_p}(\mathbf{x}) = 2\ln\left(\frac{p_X(\mathbf{x};\hat{\boldsymbol{\theta}}_p)}{p_X(\mathbf{x};\mathbf{0})}\right) = \frac{\sum_{n=0}^{p-1} x^2[n]}{\sigma^2}$. Here $\hat{\boldsymbol{\theta}}_p$ is the MLE for $\boldsymbol{\theta}_p = \{s[0], s[1], \ldots, s[p-1]\}$. We can see in Figure 3 that the MSD outperforms the EEF and MDL for all SNR.

4.2 An Application to Classification

To use the MSD for this classification problem, we first need to generalize Theorem 1 to the next corollary.

Corollary 1. *If we change the constraint set to* $\Theta = \{\boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\mathbf{C}_t^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \leq \xi^2\}$ *with* $\boldsymbol{\theta}^*$ *known, then the solution of the minimax problem*

$$\min_{\boldsymbol{\mu},\mathbf{C}} \max_{\boldsymbol{\theta}\in\Theta} D(\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu}, \mathbf{C}))$$

is $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $\mathbf{C}^* = \left(1 + \frac{\xi^2}{p}\right)\mathbf{C}_t$.

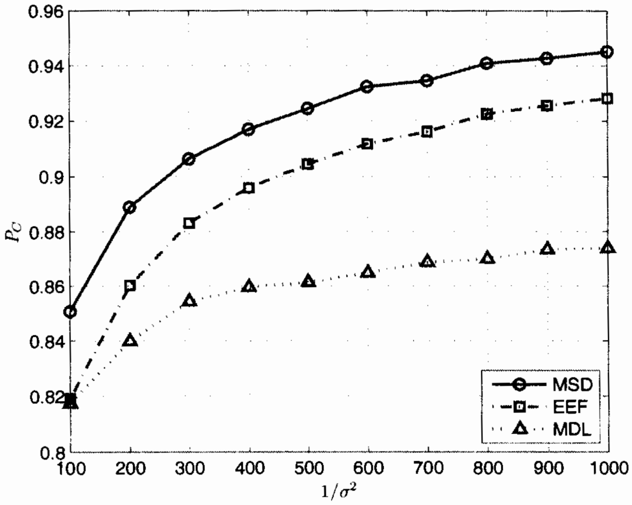


Figure 3: Performance of the MSD, EEF, and MDL for estimation of signal duration when ξ_i^2's are known.

Note that the constraint $\Theta = \{\boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{C}_t^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \leq \xi^2\}$ also makes practical sense. For example, in radar systems, we may have some prior knowledge that the possible targets are within a certain region.

Now, let us consider a classification problem where we have two models (it easily extends to multiple-model case):

$$\begin{aligned} \mathcal{M}_1 &: \mathbf{x} = \mathbf{s}_1 + \mathbf{w} \\ \mathcal{M}_2 &: \mathbf{x} = \mathbf{s}_2 + \mathbf{w} \end{aligned}$$

where $\mathbf{x}$ is $N \times 1$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ with σ^2 known, and $\mathbf{s}_1$ and $\mathbf{s}_2$ are unknown but with some partial knowledge that they satisfy $||\mathbf{s}_1 - \mathbf{s}_1^*||^2/\sigma^2 \leq \xi_1^2$ and $||\mathbf{s}_2 - \mathbf{s}_2^*||^2/\sigma^2 \leq \xi_2^2$. $\mathbf{s}_1^*$, $\mathbf{s}_2^*$, ξ_1^2 and ξ_2^2 are assumed to be known. Let $S_1 = \{\mathbf{s}_1 : ||\mathbf{s}_1 - \mathbf{s}_1^*||^2/\sigma^2 \leq \xi_1^2\}$ and $S_2 = \{\mathbf{s}_2 : ||\mathbf{s}_2 - \mathbf{s}_2^*||^2/\sigma^2 \leq \xi_2^2\}$. We also assume that S_1 and S_2 do not overlap. It follows from Corollary 1 that the minimax approximating PDFs of $\mathcal{N}(\mathbf{s}_1, \sigma^2\mathbf{I})$ and $\mathcal{N}(\mathbf{s}_2, \sigma^2\mathbf{I})$ are $\hat{p}_1(\mathbf{x}) = \mathcal{N}(\mathbf{s}_1^*, (1 + \xi_1^2/N)\sigma^2\mathbf{I})$ and $\hat{p}_2(\mathbf{x}) = \mathcal{N}(\mathbf{s}_2^*, (1 + \xi_2^2/N)\sigma^2\mathbf{I})$, respectively. As a result, we decide $\mathcal{M}_1$ if

$$\ln \hat{p}_1(\mathbf{x}) > \ln \hat{p}_2(\mathbf{x})$$

or equivalently if

$$-N\ln(1+\xi_1^2/N) - \frac{(\mathbf{x}-\mathbf{s}_1^*)^T(\mathbf{x}-\mathbf{s}_1^*)}{(1+\xi_1^2/N)\sigma^2} > -N\ln(1+\xi_2^2/N) - \frac{(\mathbf{x}-\mathbf{s}_2^*)^T(\mathbf{x}-\mathbf{s}_2^*)}{(1+\xi_2^2/N)\sigma^2}$$

Note that if $\xi_1^2 = \xi_2^2$, it is the usual ML rule for $\mathbf{s}_i = \mathbf{s}_i^*$. In order for comparison, we next derive the pseudo-ML rule for this case. The pseudo-ML rule decides $\mathcal{M}_1$ if

$$\max_{\mathbf{s}_1 \in S_1} p(\mathbf{x}; \mathbf{s}_1|\mathcal{M}_1) > \max_{\mathbf{s}_2 \in S_2} p(\mathbf{x}; \mathbf{s}_2|\mathcal{M}_2)$$

or equivalently

$$||\mathbf{x} - \hat{\mathbf{s}}_1|| < ||\mathbf{x} - \hat{\mathbf{s}}_2||$$

Note that $\hat{\mathbf{s}}_i$ minimizes the norm $||\mathbf{x}-\mathbf{s}_i||$. Therefore, if $||\mathbf{x}-\mathbf{s}_i^*||^2/\sigma^2 \le \xi_i^2$ for $i = 1$ or 2 (since S_1 and S_2 do not overlap), then $\hat{\mathbf{s}}_i = \mathbf{x}$, $||\mathbf{x}-\hat{\mathbf{s}}_i|| = 0$, and the pseudo-ML rule decides $\mathcal{M}_i$ if $\mathbf{x}$ is within the ith ball. If $||\mathbf{x} - \mathbf{s}_i^*||^2/\sigma^2 > \xi_i^2$ for $i = 1, 2$, $\hat{\mathbf{s}}_i$ must be at the intersection of the line connecting $\mathbf{x}$ and $\mathbf{s}_i^*$ and the boundary of the ball S_i for $i = 1, 2$ as illustrated in Figure 4. Therefore,

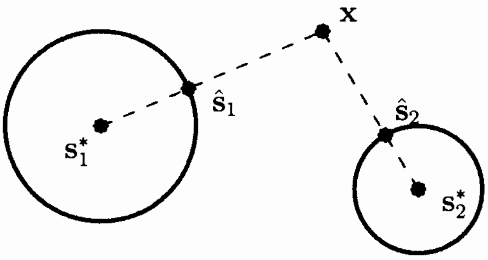


Figure 4: Solution of $\hat{\mathbf{s}}_i$ if $||\mathbf{x}-\mathbf{s}_i^*||^2/\sigma^2 > \xi_i^2$ (x is outside the ball) for $i = 1, 2$.

$||\mathbf{x} - \hat{\mathbf{s}}_i|| = ||\mathbf{x} - \mathbf{s}_i^*|| - \xi_i\sigma$ for $i = 1, 2$. The pseudo-ML rule decides $\mathcal{M}_1$ if

$$||\mathbf{x} - \mathbf{s}_1^*|| - \xi_1\sigma < ||\mathbf{x} - \mathbf{s}_2^*|| - \xi_2\sigma$$

Finally, we simulate the classification problem considered in this subsection. Let $N = 10$, $\mathbf{s}_1^* = [1\ 1\ \ldots\ 1]^T$, $\mathbf{s}_2^* = [-1\ -1\ \ldots\ -1]^T$, $\xi_1 = 0.8$, $\xi_2 = 1.2$, and $\mathcal{M}_1$ and $\mathcal{M}_2$ have equal prior probability. We plot the probability of correct classification P_C versus σ^2. For each σ^2, the experiment is repeated 50000 times. For each run, $\mathbf{s}_1$ or $\mathbf{s}_2$ are uniform randomly distributed within $\{\mathbf{s}_1 : ||\mathbf{s}_1 - \mathbf{s}_1^*||^2/\sigma^2 \leq \xi_1^2\}$ and $\{\mathbf{s}_2 : ||\mathbf{s}_2 - \mathbf{s}_2^*||^2/\sigma^2 \leq \xi_2^2\}$, respectively. In Figure 5, we can see that the MSD outperforms the pseudo-ML rule, especially when σ^2 is large.

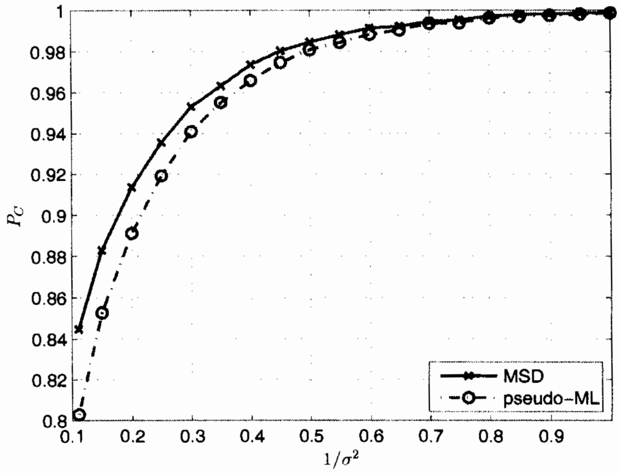


Figure 5: Classification performance of the MSD and the pseudo-ML rule.

5 The General Pythagorean Theorem for the EEF

With one sensor $T_1(\mathbf{x})$, we construct the EEF as

$$p_{\eta_1}(\mathbf{x}) = \exp\left[\eta_1 T_1(\mathbf{x}) - K^{(1)}(\eta_1) + \ln p_0(\mathbf{x})\right]$$

The true PDF $p_t(\mathbf{x})$ is unknown, but the moments $E_t\left(T_i(x)\right) = \lambda_i$ are known. In order to find η_1, we minimize $D\left(p_t || p_{\eta_1}\right) = \eta_1 E_t\left(T_1(\mathbf{x})\right) - K^{(1)}(\eta_1)$ which is equivalent to solving $E_t\left(T_1(\mathbf{x})\right) = E_{\eta_1}\left(T_1(\mathbf{x})\right)$. Let the solution be $\eta_1^{(1)*}$, and it satisfies

$$E_t\left(T_1(\mathbf{x})\right) = E_{\eta_1^{(1)*}}\left(T_1(\mathbf{x})\right) = \lambda_1 \tag{28}$$

Next, with two sensors $T_1(x)$ and $T_2(x)$, the EEF is

$$p^{(2)}_{\eta_1,\eta_2}(\mathbf{x}) = \exp\left[\eta_1 T_1(\mathbf{x}) + \eta_2 T_2(\mathbf{x}) - K^{(2)}(\eta_1, \eta_2) + \ln p_0(\mathbf{x})\right]$$

Let the solution be $\eta_1^{(2)*}$ and $\eta_2^{(2)*}$, and similarly, they satisfy

$$E_t\left(T_1(\mathbf{x})\right) = E_{\eta_1^{(2)*},\eta_2^{(2)*}}\left(T_1(\mathbf{x})\right) = \lambda_1 \tag{29}$$

$$E_t\left(T_2(\mathbf{x})\right) = E_{\eta_1^{(2)*},\eta_2^{(2)*}}\left(T_2(\mathbf{x})\right) = \lambda_2 \tag{30}$$

Then we have the following theorem:

Theorem 3 (General Pythagorean theorem for the EEF.).

$$D\left(p_t||p_0\right) = D\left(p_t||p_{\eta_1^{(2)*},\eta_2^{(2)*}}\right) + D\left(p_{\eta_1^{(2)*},\eta_2^{(2)*}}||p_{\eta_1^{(1)*}}\right) + D\left(p_{\eta_1^{(1)*}}||p_0\right) \tag{31}$$

Proof. *See Appendix B.*

6 Sensor Selection for a Multipath Scenario

Assume that we transmit a signal $s[n]$ for $0 \leq n \leq N-1$. Due to the multipath propagation, the received signal is modeled as

$$x[n] = s[n] * h[n] + w[n] \tag{32}$$

where $h[n]$ is the impulse response of the multipath model, and $w[n] \sim \mathcal{N}(0, \sigma^2)$ is white Gaussian noise. Here we assume that $h[n]$ is nonzero for $0 \leq n \leq M-1$, and therefore, we collect the received signal $x[n]$ for $0 \leq n \leq N+M-2$.

Let $\mathbf{s}_i = \left[\underbrace{0 \ \ldots \ 0}_{i \text{ 0's}} s[0] \ \ldots \ s[N-1] \ 0 \ldots 0\right]^T$ be an $(N+M-1) \times 1$ vector for $i = 0, M-1$. Then the received signal model in (32) can be written as

$$\mathbf{x} = \sum_{i=1}^{M-1} h[i]\mathbf{s}_i + \mathbf{w} = \underbrace{\left[\begin{array}{cccc} \mathbf{s}_0 & \mathbf{s}_1 & \cdots & \mathbf{s}_{M-1} \end{array}\right]}_{\mathbf{S}} \underbrace{\left[\begin{array}{c} h[0] \\ h[1] \\ \vdots \\ h[M-1] \end{array}\right]}_{\mathbf{h}} + \mathbf{w}$$

Note that $\mathbf{Sh}$ is a linear combination of $\mathbf{s}_i$'s.

Assume that we have M sensors whose outputs are

$$T_i(\mathbf{x}) = \mathbf{s}_i^T \mathbf{x}$$

for $i = 0, 1, \ldots, M-1$. We also assume that the means of the sensor outputs are known, i.e., $E\left[T_i(\mathbf{x})\right] = \mathbf{s}_i^T \mathbf{S}\mathbf{h}$ is known for $i = 0, 1, \ldots, M-1$.

Procedure:
In step 1, we choose the sensor with the smallest KL divergence $D(p_t || p_{\eta^{(1)}})$ or equivalently the longest projection of $\mathbf{Sh}$ onto the subspace generated by $\mathbf{s}_i$. Denote this sensor as $T^{(1)}(\mathbf{x})$ and corresponding vector as $\mathbf{s}^{(1)}$.

In step 2, we choose from the remaining sensors with the smallest KL divergence $D(p_t || p_{\eta^{(2)}})$ or equivalently the longest projection of $\mathbf{Sh}$ onto the subspace generated by $\left\{\mathbf{s}^{(1)}, \mathbf{s}_i\right\}$. Denote this sensor as $T^{(2)}(\mathbf{x})$ and corresponding vector as $\mathbf{s}^{(2)}$. The selected set of sensors is denoted as $\mathbf{T}^{(2)} = \left\{T^{(1)}(\mathbf{x}), T^{(2)}(\mathbf{x})\right\}$, and the corresponding set of vectors is denoted as $\mathbf{S}^{(2)} = \left\{\mathbf{s}^{(1)}, \mathbf{s}^{(2)}\right\}$.

In step i, we choose from the remaining sensors with the smallest KL divergence $D(p_t || p_{\eta^{(i)}})$ or equivalently the longest projection of $\mathbf{Sh}$ onto the subspace generated by $\left\{\mathbf{S}^{(i-1)}, \mathbf{s}_i\right\}$, and so on.

Simulation Results:
Let $s[n] = \cos(2\pi(f_0 n + \frac{1}{2}kn^2))$ for $0 \leq n \leq N-1$ where $f_0 = 0.1$, $k = 0.002$, and $N = 100$. Let $M = 20$ and the impulse response of the multipath model is plotted in Figure 6. Note that $h[3]$, $h[9]$, and $h[14]$ has the largest magnitudes,

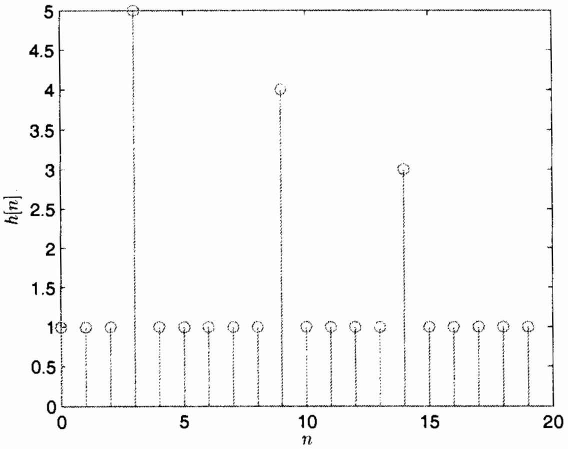


Figure 6: Impulse response of the multipath model.

and therefore, we expect to select $\mathbf{s}_3, \mathbf{s}_9, \mathbf{s}_{14}$ in the first 3 steps. The resulting order of the selected sensors is

$$3, 9, 14, 1, 18, 0, 19, 16, 5, 7, 2, 12, 10, 4, 17, 15, 8, 6, 13, 11$$

We plot the KL divergence between the true PDF and the EEF in the ith step in Figure 7. We can see that the KL divergence is close to zero at step 10, which implies that we may only need the first 10 of selected sensors to have the same performance as using all sensors.

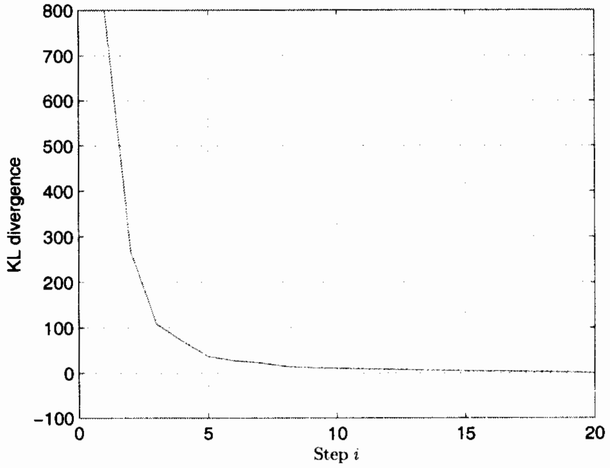


Figure 7: KL divergence between the true PDF and the EEF in the ith step.

7 Decision Combination with Multiple Learning Models/Hypotheses

We consider the combination methods that are based on estimates of posterior probabilities. Consider a training data set $\boldsymbol{D_{tr}}$ with m instances, which can be represented as $\{\boldsymbol{x_q}, y_q\}$, $q = 1, \ldots, m$, where $\boldsymbol{x_q}$ is an instance in the n dimensional feature space $\boldsymbol{X}$, and $y_q \in \boldsymbol{Y} = \{1, \ldots, C\}$ is the class identity label associated with $\boldsymbol{x_q}$. Through a training procedure, such as bootstrap sampling or subspace methods, one can develop L classifiers, h_j, $j = 1, \ldots, L$. Therefore, for each testing instance $\boldsymbol{x_t}$ in the testing data set $\boldsymbol{D_{te}}$, each classifier can vote an estimate of the *posterior* probability across all the possible class labels, $P_j(Y_i|\boldsymbol{x_t})$, $j = 1, \ldots, L$ and $Y_i = 1, \ldots, C$. Based on the Bayesian theory, given the measurements $P_j(Y_i|\boldsymbol{x_t})$, where $j = 1, \ldots, L$ and $Y_i = 1, \ldots, C$, the testing instance $\boldsymbol{x_t}$ is assigned to Y_i provided that the posterior probability is maximum. The Bayesian decision rule illustrates that it is critical to compute the probabilities of various classifiers with the consideration of all measurements simultaneously in order to fully utilize all available information to reach a prediction. The most commonly adopted combining rules in the literature include geometric average rule (GA-rule), arithmetic average rule (AA-rule), median value rule (MV-rule), majority voting rule (MajV-rule), Borda count rule (BC-rule), max and min rule, weighted average rule (Weighted AA-rule) and weighted majority voting rule (Weighted MajV-rule). The objective of our research on this is to find a combining rule for an improved estimation of the final *posterior* probability, $P(Y_i|\boldsymbol{x_t})$, based on the individual $P_j(Y_i|\boldsymbol{x_t})$ from each classifier h_j.

In this work, we consider a general ensemble learning scenario as illustrated in Fig. 8. We consider the ensemble system including L hypotheses (i.e., classifiers in our current work), each associated with a *signal strength* s_j as a criterion related to the *posterior* probability $P_j(Y_i|\boldsymbol{x_t})$. We also introduce a related concept, uncertainty degree n_j, $j = 1, \ldots, L$. For instance, in a two-class classification problem, $P_j = 0.5$ represents the lowest certainty, meaning that out of the two classes each one is equally likely. On the other hand, $P_j = 0$ or $P_j = 1$ represents full certainty, meaning that the hypothesis is certain about the class identity label. In multiclass classification problems, given a class label Y_i, the predicted label y_t of any testing instance $\boldsymbol{x_t}$ can be represented as a Boolean type: $y_t = Y_i$ or $y_t \in \bar{\boldsymbol{Y}}_i$, where $\bar{\boldsymbol{Y}}_i = \{Y_k, k \neq i\}$. In this way, the multiclass classification problem can also be transformed analogous to a two-class problem. To this end, the signal strength can be represented as $|P_j - 0.5|$, whereas the uncertainty degree is $0.5 - |P_j - 0.5|$.

The signal strength s_j and the uncertainty degree n_j can be used to represent the knowledge level of the hypothesis j. In our approach, higher weights are assigned to the classifiers that have higher signal strengths and lower uncertainty degrees, *i.e.*, are more certain on their decisions, whereas lower weights are assigned to those classifiers that have lower signal strengths and higher uncertainty degrees, *i.e.*, are less certain on their decisions. To do so, the weights ω_j should be proportional to the signal strength to uncertainty degree ratio β_j as

$$\omega_j \propto \beta_j = \frac{s_j}{n_j}, \tag{33}$$

where $s_j = |P_j - 0.5|$, and $n_j = 0.5 - s_j$. This provides the foundation of our proposed SSC approach.

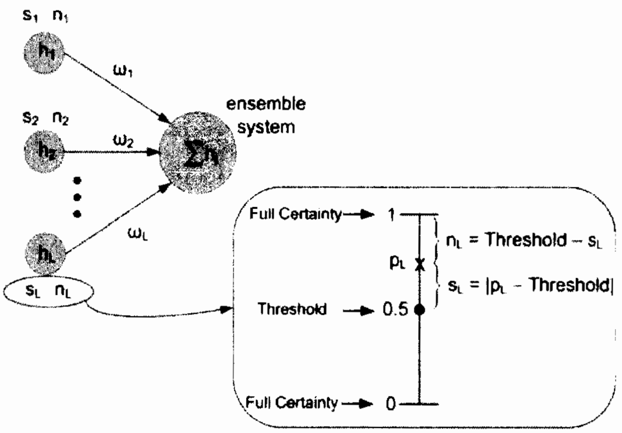


Figure 8: An ensemble learning system with multiple classifiers.

Fig. 9 gives an example of an ensemble system with four classifiers for different combining rules. Assume that we have obtained the weight coefficients (for weighted AA rule and weighted MajV rule) and the decision profile for the testing example $\boldsymbol{x_t}$. From Fig. 9 one can see, in this particular example, the MajV-rule and weighted MajV-rule vote this testing instance, $\boldsymbol{x_t}$, as a class 2 label. For the MV-rule, because the votes for class 1 and 2 are the same, the final predicted label can be randomly selected from these two classes. For the BC rule, the final predicted label can be randomly selected from classes 1, 2, and 3. All other methods vote this testing instance as a class 1 label.

To analyze the characteristics of our approach, we adopted the margin analysis to investigate the classifier decision-making process.

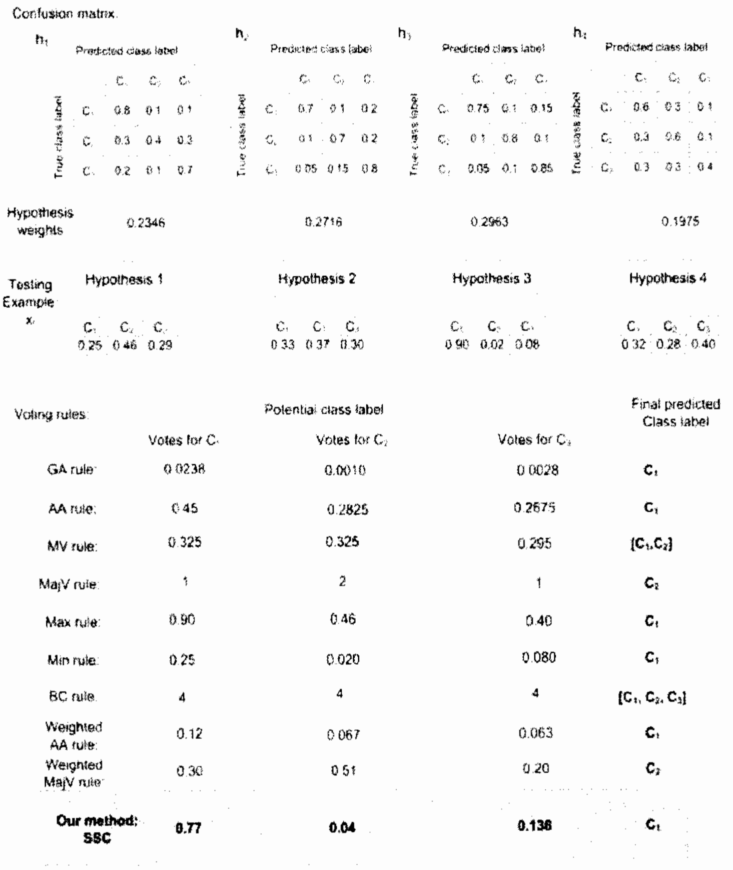
Definition 1: Consider a classification problem, the classification margin on an instance is the difference between the weight assigned to the correct label and the maximal weight assigned to any single incorrect label, *i.e.*, for an instance $\{\boldsymbol{x}, y\}$,

$$margin(\boldsymbol{x}) = w_{h(\boldsymbol{x})=y} - \max\{w_{h(\boldsymbol{x})\neq y}\} \tag{34}$$

Definition 2: Given a data distribution D, the margin distribution graph is defined as the fraction of instances whose margin is at most λ as a function of $\lambda \in [-1, 1]$:

$$F(\lambda) = \frac{|D_\lambda|}{|D|}, \lambda \in [-1, 1] \tag{35}$$

Approved for public release; distribution unlimited.



Confusion matrix:

h_1

True class label	Predicted class label: C_1	C_2	C_3
C_1	0.8	0.1	0.1
C_2	0.3	0.4	0.3
C_3	0.2	0.1	0.7

h_2

True class label	Predicted class label: C_1	C_2	C_3
C_1	0.7	0.1	0.2
C_2	0.1	0.7	0.2
C_3	0.05	0.15	0.8

h_3

True class label	Predicted class label: C_1	C_2	C_3
C_1	0.75	0.1	0.15
C_2	0.1	0.8	0.1
C_3	0.05	0.1	0.85

h_4

True class label	Predicted class label: C_1	C_2	C_3
C_1	0.6	0.3	0.1
C_2	0.3	0.6	0.1
C_3	0.3	0.3	0.4

	h_1	h_2	h_3	h_4
Hypothesis weights	0.2346	0.2716	0.2963	0.1975

Testing Example x_i

	C_1	C_2	C_3
Hypothesis 1	0.25	0.46	0.29
Hypothesis 2	0.33	0.37	0.30
Hypothesis 3	0.90	0.02	0.08
Hypothesis 4	0.32	0.28	0.40

Voting rules:	Potential class label Votes for C_1	Votes for C_2	Votes for C_3	Final predicted Class label
GA rule:	0.0238	0.0010	0.0028	**C_1**
AA rule:	0.45	0.2825	0.2675	**C_1**
MV rule:	0.325	0.325	0.295	**$[C_1, C_2]$**
MajV rule:	1	2	1	**C_2**
Max rule:	0.90	0.46	0.40	**C_1**
Min rule:	0.25	0.020	0.080	**C_1**
BC rule:	4	4	4	**$[C_1, C_2, C_3]$**
Weighted AA rule:	0.12	0.067	0.063	**C_1**
Weighted MajV rule:	0.30	0.51	0.20	**C_2**
Our method: SSC	**0.77**	**0.04**	**0.136**	**C_1**

Figure 9: Exemplary comparison of the SSC and other combining methods.

where $D_\lambda = \{\boldsymbol{x} : margin(\boldsymbol{x}) \leq \lambda\}$, $| \bullet |$ stands for the size operation and $F(\lambda) \in [0, 1]$.

Based on this, we conducted the margin analysis for the SSC method in comparison with several other method. Fig. 10 shows an example of the margin distribution graph for the proposed method with respect to the AA-rule and MV-rule. It is clearly shown that the proposed SSC method can achieve a high margin in this case. For instance, as illustrated by the dash-dotted line, for the proposed method, there are 72.06% of the testing data with margin less than 0.5, whereas for the MV-rule and AA-rule, there are 97.47% and 99.64% of the testing data with a margin less than 0.5, respectively.

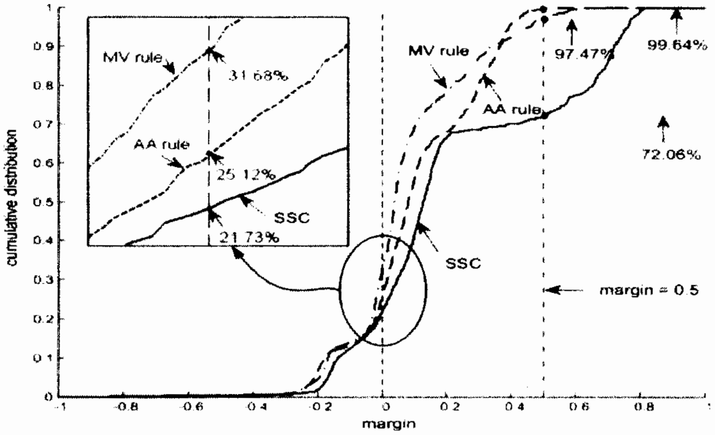


Figure 10: Margin distribution graph analysis.

8 PDF Estimation over Data sStream by Sequences of Self-Organizing Maps (SOM)

The main idea of the work is to build a series of SOMs over the data streams via two operations, *i.e.*, creating and merging the SOM sequences. The creation phase produces SOM sequence entries for windows of the data, which obtains clustering information of the incoming data streams. The size of the SOM sequences can be further reduced by combining the consecutive entries in the sequence based on the measure of Kullback-Leibler divergence. Finally, the probability density functions (pdfs) over arbitrary time periods along the data streams can be estimated using such SOM sequences. We have developed the system-level architecture, learning and estimation algorithm, and simulated the approach in Matlab. We have also compared our approach with two other KDE methods for data streams, the M-kernel approach and the cluster kernel approach, in terms of accuracy and processing time for various stationary data streams, including non-stationary data streams.

The system level architecture is illustrated in the Fig. 11.

Given a d-dimensional data, $\boldsymbol{x}_i = (x_i^1, x_i^2, \ldots, x_i^d) \in \boldsymbol{X}_I \subset \Re^d$, each neuron $\boldsymbol{n}_i$ in the SOM is associated with a d-dimensional feature vector or weight, $\boldsymbol{\omega}_i = (\omega_i^1, \omega_i^2, \ldots, \omega_i^d)$. SOM learning include three states:

Competition stage: determining the winning neuron as the neuron $\boldsymbol{n}_w(t)$

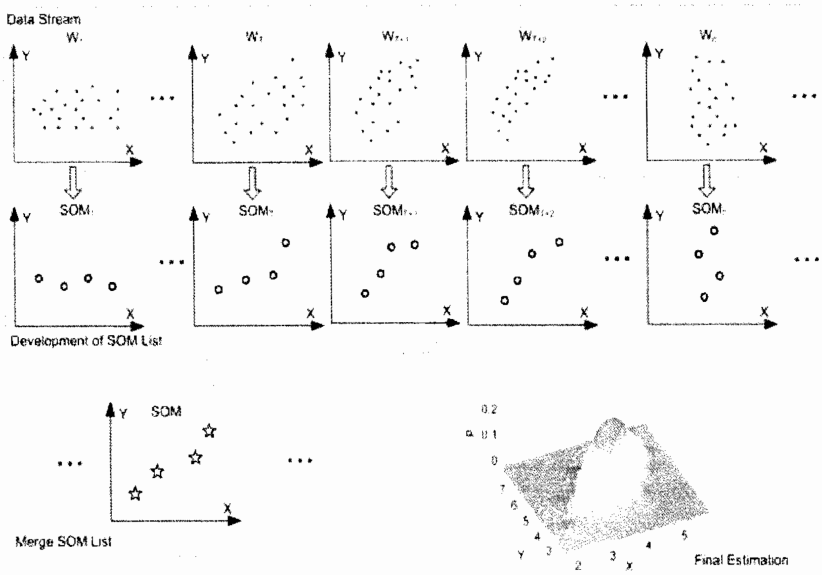


Figure 11: SOMKE: system architecture of the SOM based KDE method.

with the smallest distance (maximum similarity) to the input vector

$$\boldsymbol{\omega}_w(t) = \arg\min_{\boldsymbol{\omega}_i} \|\boldsymbol{x}_t - \boldsymbol{\omega}_i\|, \; i = 1, 2, \ldots, \ell \tag{36}$$

Cooperation stage: a topological neighborhood function around the winning neuron. Example: Gaussian function $h_i(t)$:

$$h_i(t) = \exp(-\frac{d_i^2}{2\sigma^2(t)}), \; i = 1, 2, \ldots, \ell \tag{37}$$

where d_i^2 is the squared distance on the grid of nodes between $\boldsymbol{n}_w$ and $\boldsymbol{n}_i$, and $\sigma(t)$ is the effective width of the topological neighborhood with the initial value σ_0.

Adaptation stage: Weight-updating rule:

$$\boldsymbol{\omega}_i(t+1) = \boldsymbol{\omega}_i(t) + \eta(t)h_i(t)(\boldsymbol{x}_t - \boldsymbol{\omega}_i(t)) \tag{38}$$

where $\eta(t)$ is the learning rate:

$$\eta(t) = \eta_0 \exp(-\frac{t}{\tau_2}), \quad t = 0, 1, 2, \ldots \tag{39}$$

where τ_2 is a time constant.

Once the SOM sequences are developed along the stream data, one can merge them based on estimated KL divergence (minimal KL divergence):

$$\hat{D}_{KL}(P||Q) = \int_{\Re^d} \hat{p}(x) \log \frac{\hat{p}(x)}{\hat{q}(x)} dx. \tag{40}$$

Fig. 12 illustrates an example of the combination of two SOMs, *i.e.*, $\boldsymbol{SOM}_j$ and $\boldsymbol{SOM}_{j+1}$ combine to $\boldsymbol{SOM}_c$.

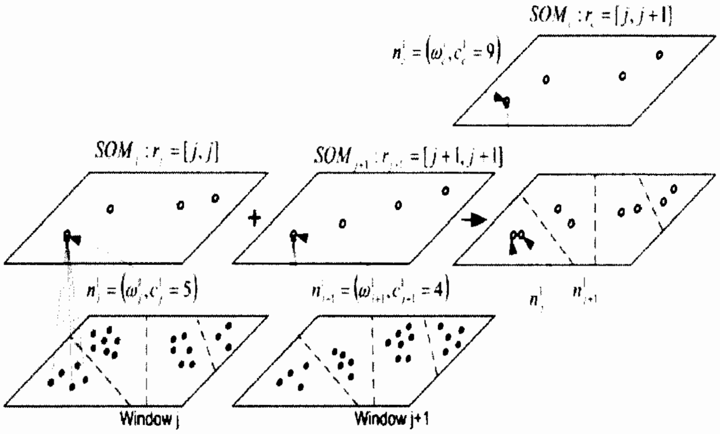

$SOM_c : r_c = [j, j+1]$
$n_c^1 = (\omega_c^1, c_c^1 = 9)$
$SOM_j : r_j = [j, j]$
$SOM_{j+1} : r_{j+1} = [j+1, j+1]$
+
$n_j^1 = (\omega_j^1, c_j^1 = 5)$
$n_{j+1}^1 = (\omega_{j+1}^1, c_{j+1}^1 = 4)$
n_j^1
n_{j+1}^1
Window j
Window j+1

Figure 12: An example of combination of two SOM sequence entries.

Two merging strategies are developed:

– Fixed Memory Strategy: Fix the total amount of memory allocated to store the SOM sequence;

– Fixed Threshold Strategy: Minimize the overall KL divergence of all consecutive pairs of entries in the SOM sequence based on a threshold α.

9 EEF for Gaussian Distribution Classification

Based on the EEF theory that the team developed, we also analyzed and developed specific classification techniques based on the EEF during this project period. For the constructed joint PDF $\hat{p}_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\eta})$, we define $l(\boldsymbol{\eta})$ as the log-likelihood function

$$l(\boldsymbol{\eta}) = \ln \hat{p}_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\eta}) \tag{41}$$

Given a set of training data, we can estimate the natural parameter vector $\hat{\boldsymbol{\eta}}$ with MLE algorithm

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} l(\boldsymbol{\eta}) \tag{42}$$

Then, we can write the constructed joint PDF of measurements under $\mathcal{H}_i$ as

$$\hat{p}_{\mathbf{T}}(\mathbf{t}, \hat{\boldsymbol{\eta}}_i) = \exp\left[< \hat{\boldsymbol{\eta}}_i, \mathbf{t} > -K_0(\hat{\boldsymbol{\eta}}_i) + \ln p_{\mathbf{T}}(\mathbf{t}, \mathcal{H}_0)\right] \tag{43}$$

Since the constructed joint PDF is parameterized by the natural parameter vector $\boldsymbol{\eta}$ in EEF, we consider the $\hat{p}_{\mathbf{T}}(\mathbf{t}, \hat{\boldsymbol{\eta}}_i)$ equal to the estimation of $p_{\mathbf{T}}(\mathbf{t}, \mathcal{H}_i)$ under $\mathcal{H}_i$.

Given unknown testing data $\mathbf{t}_s$ to be classified, similar to the MAP estimator, we assign the class $\mathcal{H}_i$ to $\mathbf{t}_s$ if $l(\boldsymbol{\eta}_i) + \ln p(\boldsymbol{\eta}_i)$ is maximized over i, where $p(\boldsymbol{\eta}_i)$ which equals to $p(\mathcal{H}_i)$ describes the prior probability of M candidate hypotheses. Under the assumption of equal prior probability of M candidate hypotheses, such that

$$p(\mathcal{H}_1) = p(\mathcal{H}_2) = \cdots = p(\mathcal{H}_M) = \frac{1}{M} \tag{44}$$

the target function of our classifier rule is built, written as

$$\ln \frac{p_{\mathbf{T}}(\mathbf{t}_s; \hat{\eta}_i)}{p_{\mathbf{T}}(\mathbf{t}_s; \mathcal{H}_0)} =< \hat{\boldsymbol{\eta}}_i, \mathbf{t}_s > -K_0(\hat{\boldsymbol{\eta}}_i) \tag{45}$$

Thus, in the training process of our new classifier rule, the natural parameters of constructed joint PDF are firstly estimated with MLE from the sets of training data under each hypothesis. Then, the constructed joint PDF for each hypothesis is available with the estimated natural parameter vectors $\hat{\boldsymbol{\eta}}$. In the testing process, the unknown testing data can be classified according to the target function built in Eq. (45), which is very similar to the MAP rule.

9.1 Hypothesis

We analyze and apply our EEF based method to Gaussian process classification. Compared to the classic expectation maximum (EM) rule for classification, our new classification rule constructs one new joint PDF $p_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\eta})$ with exponential form based on EEF and builds new target function for classification.

Considering the following hypotheses for Gaussian process classification with unknown expected mean vector and unknown covariance matrix:

$$\mathcal{H}_0 : \qquad \mathbf{x} = \mathbf{s}_0 + \mathbf{w}_0 \tag{46}$$

$$\mathcal{H}_i : \qquad \mathbf{x} = \mathbf{s}_i + \mathbf{w}_0 \tag{47}$$

where the $\mathcal{H}_0$ is the reference hypothesis and w_0 denotes Gaussian noise i.e. $\mathbf{w}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$. The data $\mathbf{s}_0$ is named as reference data which satisfies $\mathbf{s}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ where $\mathbf{I}$ is identity matrix. The meaning of $\mathbf{s}_0$ is shown later in this section. The joint PDFs of M candidate hypotheses satisfy Gaussian models i.e. we have $\mathbf{s}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i), i = 1, 2, \cdots, M$ where both $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ are unknown. The vector of $\mathbf{x}$ denoted as $[x_1, \ldots, x_p]^T$ is one data with p features or attributes. $\mathbf{s}_0$, $\mathbf{s}_i$ and $\mathbf{w}_0$ are all $p \times 1$ vectors.

In our approach, the new joint PDF with EEF is constructed based on the joint PDF under reference hypothesis. Here, we firstly assume that the covariance matrix of noise $\mathbf{\Sigma}_0$ is reasonable measured and then the joint PDF of noise could be known. To simplify our derivation, we also assume that the Gaussian process is one i.i.d. process, and that all the feature variables are conditional independent with each other. By introducing the reference data $\mathbf{s}_0$, we have $\mathcal{H}_i = \mathcal{H}_0$ by choosing $\boldsymbol{\mu}_i = \mathbf{0}$ and $\mathbf{\Sigma}_i = \mathbf{I}$. Hence, all distributions of hypotheses could be embedded into one exponential family distribution, that is the underlying idea of EEF.

9.2 Joint PDF Construction with EEF

For multivariate Gaussian process distribution, since the mean vector and covariance matrix are unknown, there is one pair of minimal sufficient statistic which are the estimation of mean vector and covariance matrix with MLE. Under the assumption that all the feature variables are conditional independent with each other, the pair of minimal sufficient statistic is written as:

$$\mathbf{T} = \mathbf{t}(\mathbf{x}) = \begin{bmatrix} \bar{\mathbf{x}} \\ \mathbf{v} \end{bmatrix} \tag{48}$$

where $\bar{\mathbf{x}}$ and $\mathbf{v}$ are the mean vector $\bar{\mathbf{x}}_k = \sum_{i=1}^{N} \mathbf{x}_i / N$ and the diagonal vector of covariance matrix $\mathbf{\Sigma} = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T / N$ respectively. Also both $\bar{\mathbf{x}}$ and $\mathbf{v}$ are the $p \times 1$ vectors, i.e. $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p]^T$ and $\mathbf{v} = [s_1^2, s_2^2, \cdots, s_p^2]^T$. N denotes the number of samples in the Gaussian process. Since all the feature variables $\mathbf{x} = [x_1, x_2, \cdots, x_p]^T$ satisfy Gaussian distribution, it is easily shown that $\bar{\mathbf{x}}, \mathbf{v}$ are independent and their joint PDFs for $\mathcal{H}_0$ are

$$\bar{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \frac{\mathbf{\Sigma}_0'}{N}) \tag{49}$$

which satisfies Gaussian distribution, and

$$p(\mathbf{v}) = \frac{\left(\frac{N}{2}\right)^{p(N-1)/2}}{\Gamma\left\{\frac{N-1}{2}\right\}^{p}} \prod_{k=1}^{p} \left(\frac{s_k}{\sigma_k^2}\right)^{N-3} \exp\left[-\frac{N}{2}\sum_{k=1}^{p}\frac{s_k^2}{\sigma_k^2}\right] \tag{50}$$

respectively, where $\mathbf{\Sigma}_0' = \mathbf{\Sigma}_0 + \mathbf{I}$ and σ_k^2 is the k-th element of diagonal matrix $\mathbf{\Sigma}_0'$.

Thus, according to the Eq. (43), the constructed joint PDF with EEF for Gaussian process is stated as:

$$\hat{p}_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\eta}) = \exp\left[< \boldsymbol{\eta}, \mathbf{t} > -K_0(\boldsymbol{\eta}) + \ln p_0(\mathbf{t})\right] \tag{51}$$

where $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^T, \boldsymbol{\eta}_2^T]^T$ in which both $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are $p \times 1$ vectors, and $\mathbf{t}(\mathbf{x})$ is shown in Eq. (48). Also, it is shown in Appendix that

$$K_0(\boldsymbol{\eta}) = \sum_{k=1}^{p} \frac{\eta_{1,k}^2 \sigma_k^2}{2N} + A_0 - \frac{N-3}{2} \sum_{k=1}^{p} \ln\left(N - 2\sigma_k^2 \eta_{2,k}\right) \tag{52}$$

where A_0 is constant term.

Therefore, the natural parameter vector $\boldsymbol{\eta}_i$ under each hypothesis $\mathcal{H}_i, i = 1, 2, \cdots, M$ can be estimated. We have

$$\begin{aligned} \frac{\partial K_0(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{\partial \boldsymbol{\eta}_1} &= \overline{\mathbf{x}} \\ \frac{\partial K_0(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{\partial \boldsymbol{\eta}_2} &= \mathbf{v} \end{aligned} \tag{53}$$

and

$$\begin{aligned} \widehat{\boldsymbol{\eta}}_{1,k} &= \frac{N\overline{x}_k}{\sigma_k^2} \\ \widehat{\boldsymbol{\eta}}_{2,k} &= \frac{N}{2\sigma_k^2} - \frac{N-3}{2s_k^2} \end{aligned} \tag{54}$$

Given one testing process $\mathbf{x}_1, \cdots, \mathbf{x}_N$ to be classified, We decide $\mathcal{H}_i$ for which the following is maximum over i:

$$\begin{aligned} \ln \frac{p_{\mathbf{T}}(\mathbf{t}, \hat{\boldsymbol{\eta}})}{p_0(\mathbf{t})} = & \sum_{k=0}^{p} \left(\widehat{\eta}_{1,k}\overline{x}_k - \frac{\widehat{\eta}_{1,k}^2 \sigma_k^2}{2N}\right) + \\ & \sum_{k=0}^{p} \left(\widehat{\eta}_{2,k} s_k^2 + \frac{N-3}{2} \ln\left(N - 2\sigma_k^2 \widehat{\eta}_{2,k}\right)\right) \end{aligned} \tag{55}$$

where the constant term is omitted.

9.3 Simulation for Gaussian Process Classification

In this part, we compare the classification performance of our rule with Pseudo-MAP for Gaussian process classification problem. Besides that, both of results are also compared with the MAP rule in which the true parameters are assumed to be known. For the models shown in Eq. (46) and Eq. (47), the Pseudo-MAP method estimates the source parameters $\boldsymbol{\mu}_i$ as $\hat{\boldsymbol{\mu}}_i$ and $\boldsymbol{\Sigma}_i$ as $\hat{\boldsymbol{\Sigma}}_i$ with MLE algorithm under each hypothesis $\mathcal{H}_i, i = 1, 2, \cdots, M$ from training data, and assumes $p_{\mathbf{T}}(\mathbf{t}, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$ is equal to the estimation of $p_{\mathbf{T}}(\mathbf{t}, \mathcal{H}_i)$. Hence, in the Pseudo-MAP method, the testing data $\mathbf{x}$ is assigned the class label when the following target function is maximum over i:

$$\begin{aligned} \ln p_{\mathbf{X}}(\mathbf{x}, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i) &= -\frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i) \\ &\quad - \frac{1}{2} \ln \left[\det \left(2\pi \hat{\boldsymbol{\Sigma}}_i \right) \right] \end{aligned} \tag{56}$$

We choose Eq. (56) and Eq. (55) as the target functions for Pseudo-MAP rule and our classifier rule respectively. There are $M = 3$ classes with $p = 5$ attributes, and $N_s = 1000$ training processes for each class. Each process has $N = 25$ samples. Let $\mathbf{s}_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ and $\omega \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, we have

$$\begin{aligned} \mu_1 &= [5, 5, 5, 5, 5] \quad \boldsymbol{\Sigma}_1 = 20\mathbf{I} \\ \mu_2 &= [5, 6, 6, 5, 5] \quad \boldsymbol{\Sigma}_2 = 21\mathbf{I} \\ \mu_3 &= [5, 5, 5, 6, 6] \quad \boldsymbol{\Sigma}_3 = 22\mathbf{I} \end{aligned} \tag{57}$$

where σ^2 is known, but the source parameters $\mu_i, \boldsymbol{\Sigma}_i, i = 1, 2, \cdots, M$ in Gaussian distributions are all unknown. The probabilities of correct classification (P_{oc}) of both rules versus σ^2 are shown in Fig. 13. Monte Carlo method is used in this simulation in which each result is averaged over 2000 experiments for every σ^2. It is shown that the classification performance of our classifier rule is very close to the performance of MAP rule. Since the construction of joint PDF EEF in our classifier rule is based on the joint PDF under the reference hypothesis and needn't the same information of training data as MAP rule, we evaluate the influence of insufficient training data for both rules. Fig. 14 shows the simulation results, in which we compare the classification performances of our classifier rule and the MAP rule versus the number of training data. It states that our rule has much better classification performance than the MAP rule as the number of training data decreased. That means, for the case of insufficient training data, our classifier rule reduces significantly the severity of decreased classification performance.

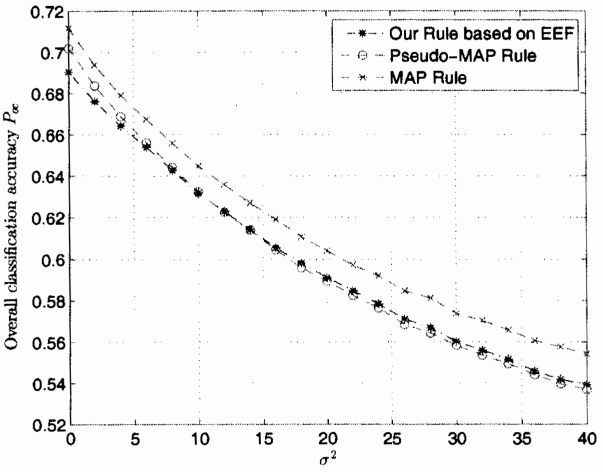


Figure 13: The probability of correct classification for our rule based on EEF and MAP rules

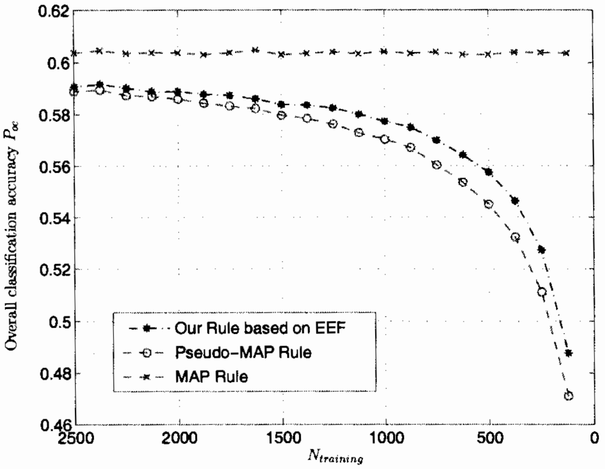


Figure 14: The impact of decreasing training data for our rule and MAP rules

10 Related Publications

1. S. Kay, Q. Ding, and M. Rangaswamy, "Sensor Integration by Joint PDF Construction using the Exponential Family," IEEE Transactions on Aerospace and Electronic Systems, vol. 49, pp. 580-593, Jan. 2013.

2. S. Kay, "A New Proof of the Neyman-Pearson Theorem using the EEF and the Vindication of Sir R. Fisher," IEEE Signal Processing Letters, vol. 19, pp. 451-454, Aug. 2012.

3. S. Kay, Q. Ding, and M. Rangaswamy, "Robust Signal Detection by Using the EEF," in Proc. the 7th IEEE Sensor Array and Multichannel Signal Processing Workshop, Jun. 2012.

3. S. Kay and Q. Ding, "Model Estimation and Classification via Model Structure Determination," IEEE Transactions on Signal Processing. (in press)

4. O. Makeyev, Q. Ding, S. Kay, and W. G. Besio, "Sensor Integration of Multiple Tripolar Concentric Ring Electrodes Improves Pentylenetetrazole-induced Seizure Onset Detection in Rats," in Proc. 34th Annual International Conference of the IEEE EMBS, Aug. 2012

5. O. Makeyev, Q. Ding, I. E. Martnez-Jurez, J. Gaitanis, S. Kay, and W. G. Besio, "Multiple Sensor Integration for Seizure Onset Detection in Human Patients Comparing Conventional Disc versus Novel Tripolar Concentric Ring Electrodes," 35th Annual International Conference of the IEEE EMBS, 2013 (under review).

6. B. Tang, H. He, Q. Ding, and S. Kay, "One Parametric Classification Rule Based on Exponentially Embedded Family with Insufficient Training Data," (in preparation)

7. H. He, S. Chen, K. Li, and X. Xu, "Incremental Learning from Stream Data," IEEE Transactions on Neural Networks, VOL. 22, NO. 12, pp. 1901-1914, December 2011.

8. H. He and Y. Cao, "SSC: A Classifier Combination Method Based on Signal Strength," IEEE Trans. Neural Networks and Learning Systems, vol. 23, issue 7, pp. 1100-1117, 2012.

9. S. Chen and H. He, "Nonstationary Stream Data Learning with Imbalanced Class Distribution", Book Chapter for Imbalanced Learning: Foundations, Algorithms, and Applications, Wiley, 2013 (in press).

10. Y. Cao, H. He, and M. Hong, "SOMKE: Kernel Density Estimation Over Data Streams by Sequences of Self-Organizing Maps," IEEE Trans. Neural Networks and Learning Systems, vol. 23, issue 8, pp. 1254-1268, 2012.

11. J. Xu, H. He, and H. Man, "DCPE co-training for classification," Neurocomputing, vol. 86, pp. 75-85, 2012.

12. Q. Cai, H. He, and H. Man, "Spatial Outlier Detection Based on Iterative Self-Organizing Learning Model," Neurocomputing, 2013. (in press)

13. Z. Ni, H. He, and J. Wen, "Adaptive Learning in Tracking Control Based on the Dual Critic Network Design," IEEE Trans. Neural Networks and Learning Systems, 2013 (in press)

14. J. Wang, H. He, D. V. Prokhorov, "A Folded Neural Network Autoencoder for Dimensionality Reduction," Procedia Computer Science, Proc. International Neural Network Society Winter Conference(INNS-WC 2012), Volume 13, pp. 120127, 2012.

15. J. Xu, Y. Yin, H. Man, and H. He, "Feature Selection Based on Sparse Imputation," in Proc. IEEE International Joint Conference on Neural Networks (IJCNN), June 10-15, 2012.

A Proof of Theorem 1

In this Appendix, we prove that $\boldsymbol{\mu}^* = \mathbf{0}$, $\mathbf{C}^* = \left(1 + \frac{\xi^2}{p}\right)\mathbf{C}_t$ is the solution of the following minimax problem:

$$\min_{\boldsymbol{\mu},\mathbf{C}} \max_{\boldsymbol{\theta}\in\Theta} D(\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu}, \mathbf{C}))$$

where the $p \times 1$ mean $\boldsymbol{\theta}$ is unknown but lies within a constraint set $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\theta}^T\mathbf{C}_t^{-1}\boldsymbol{\theta} \leq \xi^2\}$, and the covariance matrix $\mathbf{C}_t$ is known.

Proof. *From (26), we have*

$$D(\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})) = \frac{1}{2}\ln\frac{|\mathbf{C}|}{|\mathbf{C}_t|} + \frac{1}{2}tr\left[\mathbf{C}_t(\mathbf{C}^{-1} - \mathbf{C}_t^{-1})\right] + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) \tag{58}$$

Assuming that $\mathbf{C}$ *and* $\mathbf{C}_t$ *are positive definite so that the eigenvalues* $\lambda_i(\mathbf{C}) > 0$ *and* $\lambda_i(\mathbf{C}_t) > 0$ *for* $i = 1, 2, \ldots, p$. *It is well known that there exists a full rank* $p \times p$ *matrix* $\mathbf{V}$ *such that*

$$\mathbf{V}^T\mathbf{C}_t\mathbf{V} = \mathbf{I} \tag{59}$$

$$\mathbf{V}^T\mathbf{C}\mathbf{V} = \mathbf{\Lambda} \tag{60}$$

where $\mathbf{\Lambda}$ *is a diagonal matrix with positive elements, and* $\mathbf{V}$ *depends on* $\mathbf{C}$ *(since* $\mathbf{C}_t$ *is known) but not* $\boldsymbol{\mu}$. *Now since* $\mathbf{C} = (\mathbf{V}^T)^{-1}\mathbf{\Lambda}\mathbf{V}^{-1}$, *it follows that* $\mathbf{C}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T$ *and therefore,*

$$\begin{aligned} D(\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})) = &-\frac{1}{2}\ln|\mathbf{C}_t| + \frac{1}{2}\ln|(\mathbf{V}^T)^{-1}\mathbf{\Lambda}\mathbf{V}^{-1}| \\ &+ \frac{1}{2}tr\left[\mathbf{C}_t(\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T - \mathbf{C}_t^{-1})\right] \\ &+ \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T(\boldsymbol{\theta} - \boldsymbol{\mu}) \end{aligned} \tag{61}$$

Note that it follows from (59) that

$$\mathbf{C}_t = (\mathbf{V}^T)^{-1}\mathbf{V}^{-1} \tag{62}$$

$$\ln|\mathbf{C}_t| = \ln|(\mathbf{V}^T)^{-1}\mathbf{V}^{-1}| \tag{63}$$

and

$$\begin{aligned} &-\frac{1}{2}\ln|\mathbf{C}_t| + \frac{1}{2}\ln|(\mathbf{V}^T)^{-1}\mathbf{\Lambda}\mathbf{V}^{-1}| \\ &= -\frac{1}{2}\ln|\mathbf{C}_t| + \frac{1}{2}\ln|\underbrace{(\mathbf{V}^T)^{-1}\mathbf{V}^{-1}}_{\mathbf{C}_t}| + \frac{1}{2}\ln|\mathbf{\Lambda}| \\ &= \frac{1}{2}\ln|\mathbf{\Lambda}| \end{aligned}$$

Also,

$$tr(\mathbf{C}_t\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T) = tr(\underbrace{\mathbf{V}^T\mathbf{C}_t\mathbf{V}}_{\mathbf{I}}\mathbf{\Lambda}^{-1}) = tr(\mathbf{\Lambda}^{-1}) \tag{64}$$

Now let $\boldsymbol{\theta}' = \mathbf{V}^T\boldsymbol{\theta}$, $\boldsymbol{\mu}' = \mathbf{V}^T\boldsymbol{\mu}$. *Note that* $\boldsymbol{\theta}'$ *and* $\boldsymbol{\mu}'$ *depend on* $\mathbf{C}$ *since* $\mathbf{V}$ *depends on* $\mathbf{C}$. *The KL divergence can be written as*

$$\begin{aligned} D(\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})) &= \frac{1}{2}\ln|\mathbf{\Lambda}| + \frac{1}{2}tr(\mathbf{\Lambda}^{-1}) - \frac{p}{2} + \frac{1}{2}(\boldsymbol{\theta}' - \boldsymbol{\mu}')^T\mathbf{\Lambda}^{-1}(\boldsymbol{\theta}' - \boldsymbol{\mu}') \\ &= \frac{1}{2}\sum_{i=1}^{p}\ln\lambda_i + \frac{1}{2}\sum_{i=1}^{p}\frac{1}{\lambda_i} - \frac{p}{2} + \frac{1}{2}\sum_{i=1}^{p}\frac{(\theta_i' - \mu_i')^2}{\lambda_i} \end{aligned} \tag{65}$$

where λ_i's are the diagonal elements of $\mathbf{\Lambda}$.

Note that $\mathbf{C}$ enters into the KL divergence via λ_i's and implicitly in $\boldsymbol{\theta}'$ and $\boldsymbol{\mu}'$. For a fixed $\mathbf{C}$, however, we can optimize over $\boldsymbol{\theta}'$ and $\boldsymbol{\mu}'$. To simplify the optimization, let $a_i = 1/\lambda_i > 0$ for $i = 1, 2, \ldots, p$. The KL divergence is

$$\begin{aligned} D(\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)||\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})) &= J(\boldsymbol{\theta}', \boldsymbol{\mu}', \mathbf{a}) \\ &= -\frac{p}{2} + \frac{1}{2}\sum_{i=1}^{p}(a_i - \ln a_i) + \frac{1}{2}\sum_{i=1}^{p} a_i(\theta_i' - \mu_i')^2 \end{aligned} \quad (66)$$

Now we determine

$$\min_{\mathbf{a}, \boldsymbol{\mu}'} \max_{\boldsymbol{\theta}'} J(\boldsymbol{\theta}', \boldsymbol{\mu}', \mathbf{a}) \quad (67)$$

subject to $\boldsymbol{\theta}^T \mathbf{C}_t^{-1} \boldsymbol{\theta} \leq \xi^2$. Since $\mathbf{C}_t^{-1} = \mathbf{V}\mathbf{V}^T$, in the $\boldsymbol{\theta}'$ space the equivalent constraint is $\boldsymbol{\theta}'^T \boldsymbol{\theta}' \leq \xi^2$.

First consider maximizing

$$g_1(\boldsymbol{\theta}') = \sum_{i=1}^{p} a_i(\theta_i' - \mu_i')^2 \quad (68)$$

over $\boldsymbol{\theta}'$. But

$$(\theta_i' - \mu_i')^2 \leq (|\theta_i'| + |\mu_i'|)^2 \quad (69)$$

with equality holds for $\theta_i' = -sgn(\mu_i')|\theta_i'|$ where $sgn(\cdot)$ is the sign function. Therefore, we can equivalently maximize

$$g_2(|\boldsymbol{\theta}'|) = \sum_{i=1}^{p} a_i(|\theta_i'| + |\mu_i'|)^2 \quad (70)$$

over $|\boldsymbol{\theta}'|$. For fixed $\mathbf{a}$, note that

$$\begin{aligned} & \min_{\boldsymbol{\mu}'} \max_{|\boldsymbol{\theta}'|} \sum_{i=1}^{p} a_i(|\theta_i'| + |\mu_i'|)^2 \\ \geq & \max_{|\boldsymbol{\theta}'|} \min_{\boldsymbol{\mu}'} \sum_{i=1}^{p} a_i(|\theta_i'| + |\mu_i'|)^2 \\ = & \max_{\boldsymbol{\theta}' : \boldsymbol{\theta}'^T \boldsymbol{\theta}' \leq \xi^2} \sum_{i=1}^{p} a_i \theta_i'^2 \\ = & \max_{\boldsymbol{\theta}' : \boldsymbol{\theta}'^T \boldsymbol{\theta}' = \xi^2} \sum_{i=1}^{p} a_i \theta_i'^2 \\ = & a_{\max} \xi^2 \end{aligned} \quad (71)$$

where $a_{\max} = \max\{a_1, a_2, \ldots, a_p\}$. *Equality holds if and only if* $\boldsymbol{\mu}' = \mathbf{0}$, $\boldsymbol{\theta}' = [0 \ \cdots \ 0 \ \xi \ 0 \ \cdots \ 0]^T$ *where the position of the only non-zero element* ξ *corresponds to that of* $a_{\max}$. *As a result, for fixed* $\mathbf{a}$,

$$\min_{\boldsymbol{\mu}'} \max_{\boldsymbol{\theta}'} J(\boldsymbol{\theta}', \boldsymbol{\mu}', \mathbf{a}) = -\frac{p}{2} + \frac{1}{2}\sum_{i=1}^{p}(a_i - \ln a_i) + \frac{1}{2}a_{\max}\xi^2 \tag{72}$$

It only remains to minimize $f(\mathbf{a}) = \sum_{i=1}^{p}(a_i - \ln a_i) + a_{\max}\xi^2$ *over* $\mathbf{a}$ *with* $a_i > 0$ *for all* i.

First note that $a_i - \ln a_i$ *is a strictly convex function in* a_i *and hence* $\sum_{i=1}^{p}(a_i - \ln a_i)$ *is strictly convex in* $\mathbf{a}$. *Also* $a_{\max} = \max\{a_1, a_2, \ldots, a_p\}$ *is convex. Thus,* $f(\mathbf{a})$ *is strictly convex and a symmetric function of the* a_i*'s. If a minimum exists, it is unique. Hence, the minimum must be at* $a_1 = a_2 = \cdots = a_p$, *since otherwise any permutation will produce the same value, violating uniqueness. Letting* $a = a_1 = a_2 = \cdots = a_p$, *the objective function is*

$$f(\mathbf{a}) = p(a - \ln a) + a\xi^2 \tag{73}$$

Taking the derivative with respect to a *and setting it to zero,* a *satisfies*

$$p - \frac{p}{a} + \xi^2 = 0$$

producing $a^* = \frac{1}{1+\xi^2/p}$. *From (72), we have*

$$\min_{\mathbf{a},\boldsymbol{\mu}'} \max_{\boldsymbol{\theta}'} J(\boldsymbol{\theta}', \boldsymbol{\mu}', \mathbf{a}) = -\frac{p}{2} + \frac{p}{2}\left(\frac{1}{1+\xi^2/p} + \ln(1+\xi^2/p)\right) + \frac{1}{2}\frac{\xi^2}{1+\xi^2/p} = \frac{p}{2}\ln(1+\xi^2/p) \tag{74}$$

which is the minimax KL divergence between $\mathcal{N}(\boldsymbol{\theta}, \mathbf{C}_t)$ *and* $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$.

Finally, we need to find the minimax solution of $\boldsymbol{\mu}$ *and* $\mathbf{C}$. *First, note that*

$$\boldsymbol{\mu}^* = (\mathbf{V}^T)^{-1}\boldsymbol{\mu}'^* = \mathbf{0} \tag{75}$$

Also, $\lambda_i^* = 1/a^* = 1 + \xi^2/p$ *for all* i *and thus* $\boldsymbol{\Lambda}^* = (1 + \xi^2/p)\mathbf{I}$. *It follows that*

$$\begin{aligned} \mathbf{C}^* &= (\mathbf{V}^T)^{-1}\boldsymbol{\Lambda}^*\mathbf{V}^{-1} \\ &= (\mathbf{V}^T)^{-1}(1+\xi^2/p)\mathbf{I}\mathbf{V}^{-1} \\ &= (1+\xi^2/p)\underbrace{(\mathbf{V}^T)^{-1}\mathbf{V}^{-1}}_{\mathbf{C}_t} \\ &= (1+\xi^2/p)\mathbf{C}_t \end{aligned} \tag{76}$$

B Proof of Theorem 3

Proof. *We have that*

$$\begin{aligned}
D\left(p_t || p_{\eta_1^{(2)*},\eta_2^{(2)*}}\right) &= E_t\left(\ln p_t(\mathbf{x}) - \left[\eta_1^{(2)*}T_1(\mathbf{x}) + \eta_2^{(2)*}T_2(\mathbf{x}) - K^{(2)}(\eta_1^{(2)*},\eta_2^{(2)*}) + \ln p_0(\mathbf{x})\right]\right) \\
&= D\left(p_t||p_0\right) - E_t\left(\eta_1^{(2)*}T_1(\mathbf{x}) + \eta_2^{(2)*}T_2(\mathbf{x}) - K^{(2)}(\eta_1^{(2)*},\eta_2^{(2)*})\right) \\
&= D\left(p_t||p_0\right) - \left(\eta_1^{(2)*}\lambda_1 + \eta_2^{(2)*}\lambda_2 - K^{(2)}(\eta_1^{(2)*},\eta_2^{(2)*})\right) \quad (77)
\end{aligned}$$

It follows from (29) and (30) that

$$\begin{aligned}
D\left(p_{\eta_1^{(2)*},\eta_2^{(2)*}}||p_{\eta_1^{(1)*}}\right) &= E_{\eta_1^{(2)*},\eta_2^{(2)*}}\left(\eta_1^{(2)*}T_1(\mathbf{x}) + \eta_2^{(2)*}T_2(\mathbf{x}) - K^{(2)}(\eta_1^{(2)*},\eta_2^{(2)*})\right) \\
&\quad - E_{\eta_1^{(2)*},\eta_2^{(2)*}}\left(\eta_1^{(1)*}T_1(\mathbf{x}) - K^{(1)}(\eta_1^{(1)*})\right) \\
&= \eta_1^{(2)*}\lambda_1 + \eta_2^{(2)*}\lambda_2 - K^{(2)}(\eta_1^{(2)*},\eta_2^{(2)*}) - \left[\eta_1^{(1)*}\lambda_1 - K^{(1)}(\eta_1^{(1)*})\right] \\
&\qquad (78)
\end{aligned}$$

From (28), we have that

$$D\left(p_{\eta_1^{(1)*}}||p_0\right) = E_{\eta_1^{(1)*}}\left(\eta_1^{(1)*}T_1(\mathbf{x}) - K^{(1)}(\eta_1^{(1)*})\right) = \eta_1^{(1)*}\lambda_1 - K^{(1)}(\eta_1^{(1)*}) \quad (79)$$

Summing (77), (78), and (79) results in the general Pythagorean theorem in (31).